

# Usefulness of ORD from Literatures Review

13-15 Oct. 2025

IWG-DDADWS #9 Workshop in Brussels

JASIC technical WG for regulation of driver monitoring system

Yohei IWASHITA



# Today's overview

- At the previous workshop, we introduced the Japanese guideline, that adopt “Observer Rating of Drowsiness”(ORD) as a standard method for measuring drowsiness.
- Therefore, the ORD has long been used in research & development on driver drowsiness and the monitoring system in Japanese industries.
- In contrast, the EU GSR designates KSS as the standard method, while ORD is considered an alternative.
- In Japan, there are some concerns that using ORD requires evidence showing its relationship with KSS, making its adoption more challenging.
- We believe ORD is a valid method for measuring drowsiness in principle with comparable to KSS. We hope to discuss treating ORD not as an “alternative,” but as “one of the standard methods” in the 01 series.

# Contents

- Validity of ORD
  - Is ORD a method that can measure what we intend to measure?
- Study of Warning Threshold
  - How can we determine the level of drowsiness in ORD as warning threshold?
- Summary and Next step

# Humans naturally have the ability to perceive facial expressions



*Created by Microsoft Copilot*

- According to Ekman's "Basic Emotion Theory" (1972), emotions are biologically universal and are linked to specific facial expressions.
- Studies using multidimensional scaling based on perceived facial similarity identified arousal level as the second factor. (Abelson & Sermat, Gladstones, Russell & Bullock)
- These studies suggest that humans have the ability to perceive facial expressions, and that many people can give the same ratings to the same facial expressions.

# Correlation between ORD and others (Uchiyama et al. (2023))

Analysis of correlations between ORD and some physiological responses, KSS, driving indicator, which are related to drowsiness

## Experiment

- Condition: Expressway at night on driving simulator
- Participants: 17

## ORD method

- ORD assessors: 3(trained)
- ORD scale: 5 levels
- Concordance rate: >0.7

Variables	1	2	3	4	5	6
1. ORD						
2. KSS	1.00 <sup>***</sup> (0.88-1.11)					
3. SDLP	1.29 <sup>***</sup> (1.15-1.43)	0.71 <sup>***</sup> (0.60-0.82)				
4. PERCLOS	1.06 <sup>***</sup> (0.78-1.34)	0.55 <sup>***</sup> (0.39-0.72)	1.07 <sup>***</sup> (0.70-1.43)			
5. Percentage of time occupied by SEM	0.98 <sup>***</sup> (0.75-1.20)	0.60 <sup>***</sup> (0.45-0.75)	1.13 <sup>***</sup> (0.81-1.44)	0.80 <sup>***</sup> (0.45-1.15)		
6. EEG alpha power	0.88 <sup>***</sup> (0.67-1.09)	0.62 <sup>***</sup> (0.41-0.84)	0.73 <sup>***</sup> (0.58-0.89)	0.61 <sup>***</sup> (0.43-0.80)	0.55 <sup>***</sup> (0.39-0.71)	
7. EEG theta power	0.52 <sup>**</sup> (0.24-0.81)	0.24 <sup>*</sup> (0.05-0.43)	0.42 <sup>**</sup> (0.18-0.66)	0.45 <sup>***</sup> (0.23-0.66)	0.30 <sup>*</sup> (0.07-0.53)	0.66 <sup>***</sup> (0.37-0.95)

フィッシャーのz変換を行った  
ピアソン相関係数

ORD, Observer Rating of Drowsiness; KSS, Karolinska Sleepiness Scale; SDLP, Standard Deviation of Lateral Position; PERCLOS, PERcentage of eye CLOSure; SEM, Slow Eye Movement; EEG, electroencepharogram; The values are mean Fisher's z-values transformed from Pearson correlation coefficients across 17 participants. The z-values were tested with one-sample t-test. Parentheses indicate 95% confidence interval.

<sup>\*</sup>,  $P < 0.05$ ;  
<sup>\*\*</sup>,  $P < 0.01$ ;  
<sup>\*\*\*</sup>,  $P < 0.001$

Source: [Yuji Uchiyama et al., Convergent validity of video-based observer rating of drowsiness, against subjective, behavioral, and physiological measures \(2023\)](#)

ORD correlates with physiological responses related to drowsiness and other indices such as KSS.

# Correlation between ORD and EEG on real road (Tiadi et al. (2024))

Self-report (KSS) and ORD were compared with EEG (electroencephalogram) as ground truth.

## Experiment

- Condition: Real road driving
  - Condition A: Baseline
  - Condition B: Deprived of sleep
- Participants: 50

## ORD method

- ORD assessors: 6(trained)
- ORD scale: 9 levels
- Concordance rate: 0.92(avg.)

### ORD

Baseline Condition A		Observer Ratings	
		Not drowsy (KSS <7)	Least drowsy KSS (=7)
EEG Outputs	Not drowsy	100.00%	0.00%
	Least drowsy	100.00%	0.00%

### KSS(self report)

Baseline Condition A		Driver Self-Reports	
		Not drowsy (KSS <7)	Least drowsy KSS (=7)
EEG Outputs	Not drowsy	100.00%	0.00%
	Least drowsy	100.00%	0.00%

Deprived of sleep Condition B		Observer Ratings	
		Least drowsy KSS (=7)	Drowsy (KSS >=8)
EEG Outputs	Least drowsy	74.51%	25.49%
	Drowsy	5.88%	94.12%

Deprived of sleep Condition B		Driver Self-Reports	
		Least drowsy KSS (=7)	Drowsy (KSS >=8)
EEG Outputs	Least drowsy	65.38%	34.62%
	Drowsy	12.50%	87.50%

Source [Tiadi et al., Exploring Driver's self-report and observer ratings of driver drowsiness based on real road driving\(2024\)](#)

They concluded that both ORD and KSS can be used to measure drowsiness.

# Drowsiness and crash with NDS data (Dingus et al.,(2016))

Using Naturalistic Driving Study data, the contributing factors to driver crash risk and their occurrence frequency are analyzed. The following are results related to “drowsiness and fatigue”.

## Method

- **Drowsiness/Fatigue** is defined as a state observable in the "20 seconds video segment just before the crash."
- Labeling drowsiness/fatigue or not with observable behaviors of driver. (not use ORD)

## Result

- **Prevalence:** Drowsiness/fatigue was observed in 1.57% of baselines.
- **Crash risk:** Drowsiness/fatigue is risky compared to normal driving condition (i.e., Odds ratio of 3.4).

There is a strong relationship between “observable drowsy signs” and crash risk, supporting the usefulness of ORD as indirect evidence.

	O.R. (95% CI)	Baseline Prevalence
<b>Observable Impairment*</b>		
Overall	5.2 (3.8 - 7.1)	1.92%
Drug/alcohol	35.9 (17.0 - 75.8)	0.08%
Drowsiness/fatigue	3.4 (2.3 - 5.1)	1.57%
Emotion (anger, sadness, crying, and/or emotional agitation)	9.8 (5.0 - 19.0)	0.22%
<b>Driver Performance Error</b>		
Overall	18.2 (14.8 - 22.3)	4.81%
Major error sub-categories (observed in crash and baseline events)		
Apparent inexperience with vehicle/roadway	204.5 (111.1 - 376.6)	0.07%
Blind spot error	55.1 (21.6 - 140.6)	0.05%
Improper turn	92.1 (68.8 - 123.4)	0.51%
Right-of-way error	936.1 (123.8 - 7078.3)	0.01%
Signal violation	28.3 (15.9 - 50.2)	0.19%
Stop/yield sign violation	7.4 (4.9 - 11.4)	1.05%
Wrong side of road	22.3 (12.0 - 41.5)	0.19%
Driving too slowly	2.3 (1.1 - 4.8)	0.97%
Sudden or improper braking/stopping	247.8 (53.1 - 1156.2)	0.01%
Failed to signal	2.5 (1.5 - 4.0)	2.27%
<b>Driver Momentary Judgment Error (Speeding/Aggressive Driving)</b>		
Overall	11.1 (9.0 - 13.8)	4.22%
Aggressive driving (general observed behavior)	34.8 (17.2 - 70.5)	0.10%
Speeding (over limit and too fast for conditions)	12.8 (10.1 - 16.2)	2.77%
Speeding/unsafe in work zone	14.2 (3.9 - 52.0)	0.05%
Illegal/unsafe passing	14.4 (7.2 - 28.8)	0.18%
Following too closely	13.5 (4.4 - 41.4)	0.07%
Intentional signal violation	15.3 (7.9 - 29.9)	0.19%
Intentional stop/yield sign violation	5.3 (3.4 - 8.4)	1.04%
<b>Observable Distraction**</b>		
Overall	2.0 (1.8 - 2.4)	51.93%
Major distraction sub-categories (observed in crash and baseline events)		
In-vehicle radio	1.9 (1.2 - 3.0)	2.21%
In-vehicle climate control	2.3 (1.1 - 5.0)	0.56%
In-vehicle device (other)	4.6 (2.9 - 7.4)	0.83%
Total in-vehicle device	2.5 (1.8 - 3.4)	3.53%
Cell browse	2.7 (1.5 - 5.1)	0.73%
Cell dial (handheld)	12.2 (5.6 - 26.4)	0.14%
Cell reach	4.8 (2.7 - 8.4)	0.58%
Cell text (handheld)	6.1 (4.5 - 8.2)	1.91%
Cell talk (handheld)	2.2 (1.6 - 3.1)	3.24%
Total cell (handheld)	3.6 (2.9 - 4.5)	6.40%
Child rear seat	0.5 (0.1 - 1.9)	0.80%
Interaction with adult/teen passenger	1.4 (1.1 - 1.8)	14.58%
Reading/writing (includes tablet)	9.9 (3.6 - 26.9)	0.09%
Eating	1.8 (1.1 - 2.9)	1.90%
Drinking (non-alcohol)	1.8 (1.0 - 3.3)	1.22%
Personal hygiene	1.4 (0.8 - 2.5)	1.69%
Reaching for object (non-cell phone)	9.1 (6.5 - 12.6)	1.08%
Dancing in seat to music	1.0 (0.4 - 2.3)	1.10%
Extended glance duration to external object	7.1 (4.8 - 10.4)	0.93%

The baseline prevalence of a factor represents the percentage of time the factor was present during the normal driving condition.

\*Observable from 20-second pre-crash and baseline sample video segments

\*\*Observable from 6-second pre-crash and baseline sample video segments

# Contents

- Validity of ORD
  - Is ORD a method that can measure what we intend to measure?
- **Study of Warning Threshold**
  - **How can we determine the level of drowsiness in ORD as warning threshold?**
- Summary and Next step

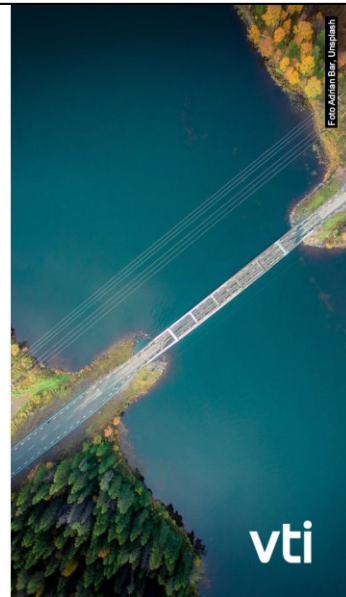
# Warning threshold for KSS

## Research on drowsiness levels that require warning

- A warning is required when the KSS is 8. This is based on research showing that the risk of an accident increases in relation to driving performance (reported by VTI at workshop #8).

### KSS in relation to driving performance

- KSS correlates with lane-keeping performance, number of line crossings and the time headway.
  - In simulator studies, line crossings start to appear at  $KSS \geq 7$  (Ingre et al. 2006)
  - Crossings with 4 wheels increased 28 times at  $KSS = 8$  and 185 times at  $KSS = 9$
  - First rumble strip hit occurred at  $KSS 8.1$  (Anund et al. 2008)
- Self-rated fitness-to-drive drops at  $KSS > 7$
- Levels 8 and 9 are critical



### KSS in relation to driving performance

- KSS is a significant predictor of line crossings with an OR of 5.4
- Clear deterioration when  $KSS \geq 8$

**Table 3** The number of line crossings during day and night-time at different KSS levels. Percentages of observations distributed per KSS, separated for day/night, in brackets

KSS	Day		Night	
	0 line crossings	$\geq 1$ line crossing	0 line crossings	$\geq 1$ line crossing
1 (KSS 1-5)	35 (100)	0	0	0
2 (KSS 6)	20 (100)	0	10 (100)	0
2 (KSS 7)	19 (95)	1 (5)	16 (100)	0
3 (KSS 8)	6 (75)	2 (25)	23 (88)	3 (12)
3 (KSS 9)	6 (67)	3 (33)	25 (60)	17(40)

Anund et al 2017

Eur. Transp. Res. Rev. (2017) 9: 31  
DOI 10.1007/s12544-017-0248-6



Source: DDADWS-08-09 (VTI) Methodological Considerations when Evaluating DDAW Systems ASD.pdf

It is concluded that  $KSS=8$  is clearly dangerous based on the number of line crossings.

# Warning threshold for ORD

What levels of ORD should be required warning?

- Under GSR and the 00 series, when using alternative measurements, evidence is required to show that the drowsiness level triggered a warning is equivalent to “KSS=8”.

## Approach to show the equivalence

- A) Based on the correspondence between KSS and ORD, the ORD level equivalent to KSS=8 is identified.
- B) Based on the relationship between ORD levels and driving performance, the threshold level requiring a warning is identified.

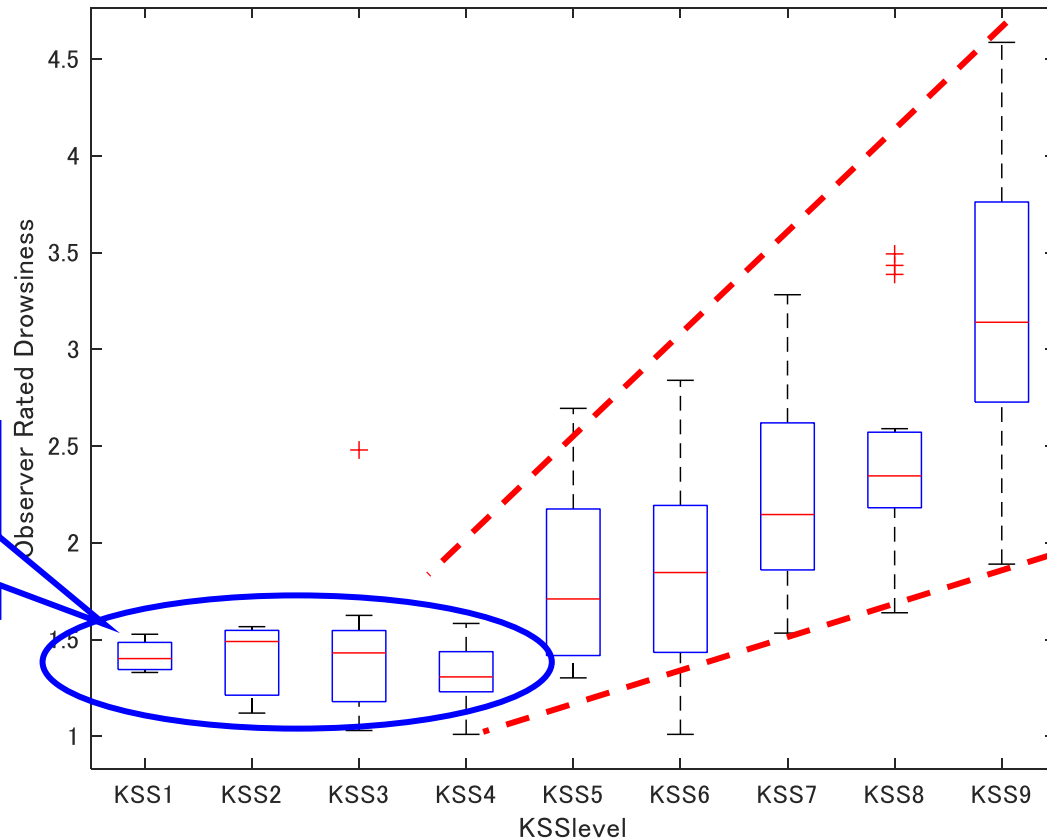
# A) Study on correspondence between KSS and ORD

- Based on the verbal descriptions of the KSS scale, scores of KSS=5 (“Neither alert nor sleepy”) and below can be considered to indicate no drowsiness and can be interpreted as equivalent to D1 in the ORD(5-point-scale).
  - D1 = KSS1 to KSS5
- Regarding the maximum drowsiness score, both scales are considered to represent the state just before sleeping and can be interpreted as indicating the same state.
  - D5 (Extremely drowsy) = KSS9 (Very sleepy, great effort to keep awake, fighting sleep)

Lv	ORD	Karolinska Sleepiness Scale (KSS)	DDAW (00 series)
D1	Not drowsy	1. Extremely alert 2. Very alert 3. Alert 4. Rather alert 5. Neither alert nor sleepy	Not mentioned
D2	Slightly drowsy	6. Some signs of sleepiness	
D3	Moderately drowsy	7. Sleepy, no effort to keep awake	(Warning)
D4	Very drowsy	8. Sleepy, some effort to keep awake	
D5	Extremely drowsy	9. Very sleepy, great effort to keep awake, fighting sleep	<b>Warning</b>

# A) Study on correspondence between KSS and ORD

The correspondence was examined using open data from Uchiyama et al.(2023).



*ORD remains almost constant low with low KSS. (as expected)*

*ORD have a wide range of variability above KSS=5.*

Although KSS and ORD are correlated, the variability makes it difficult to indicate precise correspondence.

# Why is so variable on ORD?

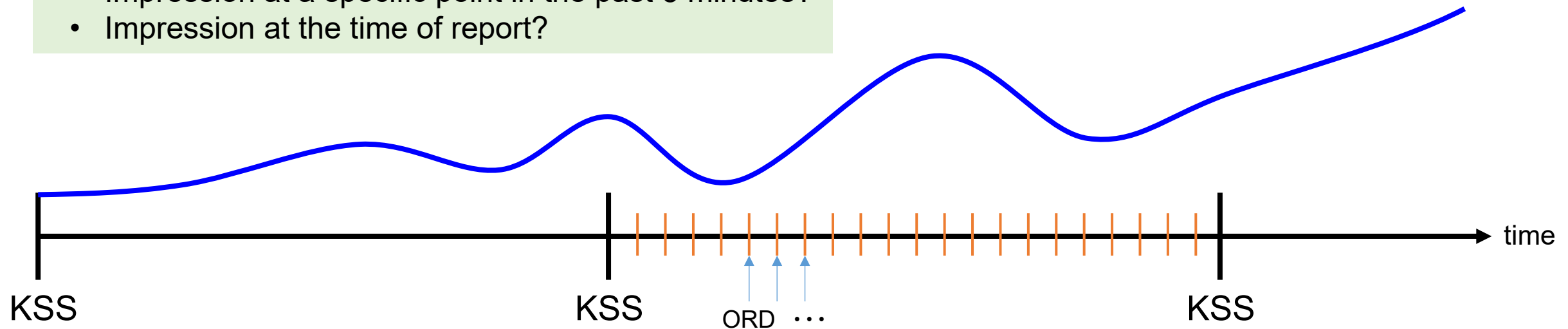
In Uchiyama et al.(2023);

- KSS is evaluated every 5 minutes
- ORD is evaluated every 5 seconds and average of 5 minutes

KSS may be evaluated differently by participants

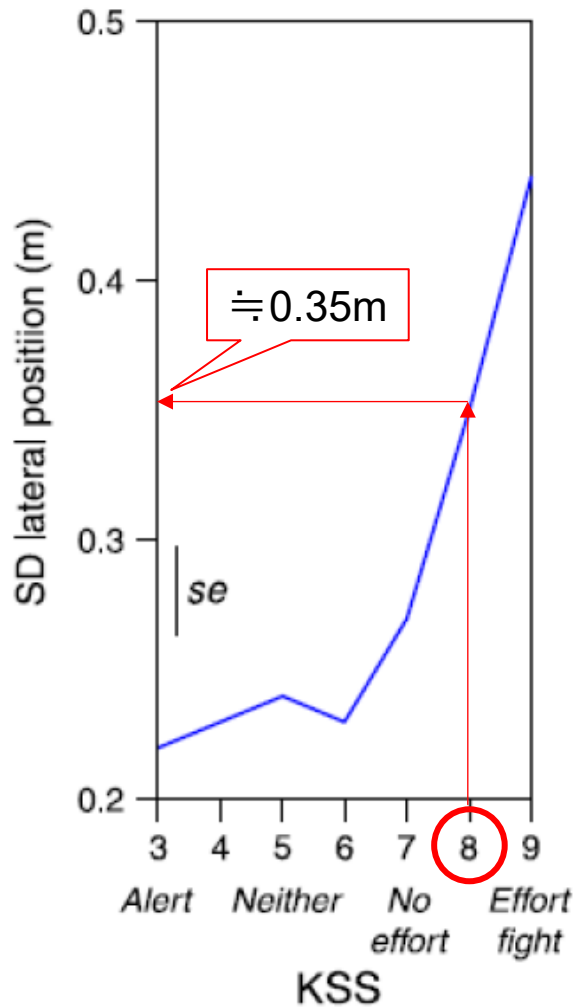
- Overall impression over the past 5 minutes?
- Impression at a specific point in the past 5 minutes?
- Impression at the time of report?

ORD is the average for the past 5 minutes (i.e., the overall rating)

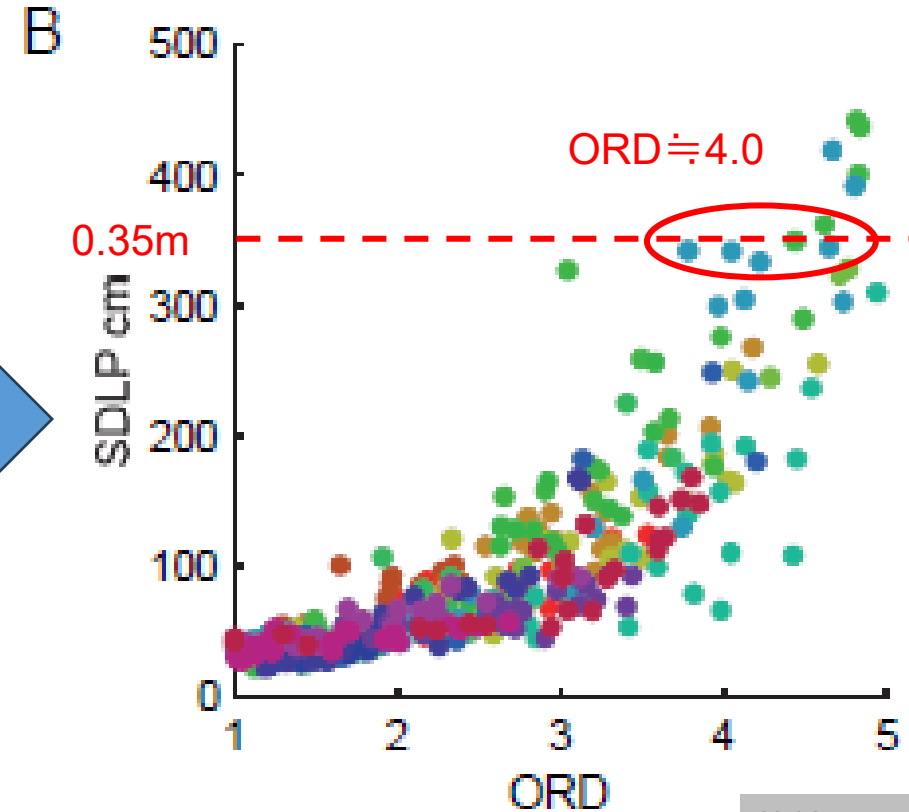


The greater the change in drowsiness over a 5-minute period, the more noticeable the difference.  
(However, there should be proper ways to suppress the difference)

## B) Study on relationship between ORD and driving performance



Very similar result



Uchiyama et al. (2023)

The SDLP corresponding to KSS=8 is 0.35m and the corresponding ORD is around 4.0.

Although more evidence is needed, this seems to be easier to indicate.

# Contents

- Validity of ORD
  - Is ORD a method that can measure what we intend to measure?
- Study of Warning Threshold
  - How can we determine the level of drowsiness in ORD as warning threshold?
- **Summary and Next step**

# Summary

- Humans have the ability to perceive others' facial expressions, and with appropriate training, anyone can assess drowsiness.
- ORD, conducted by multiple trained assessors, is a valid method for measuring drowsiness, as it has been shown to correlations with other drowsiness indicators.
- It may be difficult to identify a specific ORD level for a warning threshold based on the correspondence between KSS and ORD.
- To identify the appropriate ORD warning threshold, it is more reasonable to base it on its relationship with driving performance metrics (e.g., SDLP, lane crossings), though further data collection is needed.

## Advantage of adding the ORD to 01 series

- Assuming that KSS and ORD can measure the same drowsiness, ORD requires more cost than KSS, such as training for assessors, but has the following advantages:
  - It can be evaluated later as long as there is video data.
  - Using common assessors helps reduce variability across measurements.
- These advantages may help enhance the fairness of tests, especially in technical services:
  - Driving test can be conducted without intervention from the experimenter while recording video of the driver, and the drowsiness can be assessed after the drive.
  - It will be possible to evaluate different systems using the same criteria using trained assessors in technical service.
- Of course, both KSS and ORD are OK for internal verification of OEMs.

# Future work for 01 series

6.2.2. Where alternative measurements to KSS are used to determine the participant's level of drowsiness, the manufacturer shall provide evidence that the chosen measurement is a valid and accurate means to assess driver drowsiness, and that the drowsiness threshold used in the validation testing is equivalent to a KSS level referred to in paragraph 5.5.2. in this Regulation.

6.2.2.1. For the sleep video analysis, expected evidence concerns the quality of the video used, the visibility of the setup for the participant, the correspondence between the rating scale and the KSS, the training of the assessors (in addition a minimal performance level of 'concordance rate' superior or equal to 0.70 is required), information of independence of the assessors to the DDAW system development, and description on how the final rating is calculated based on the input from the sleep experts.

## 7. Equivalence between alternative drowsiness measurements and KSS

7.1. If alternative measurements to KSS are used to validate a DDAW system, the manufacturer shall state the threshold being used and provide evidence detailing the equivalency between the threshold and a KSS level of 8.

7.1.1. If the alternative measurement uses a scale which has fewer descriptive levels than KSS, the equivalence between the alternative scale and KSS shall refer to the lowest corresponding level when compared to KSS.

- If you agree with treating ORD the same as KSS, we would like to bring a revised proposal for the 01 series to the next IWG meeting.
- We would like to consider and propose the followings:
  - How to explain the warning threshold.
  - Necessity of evidence about the correspondence with KSS
  - etc.

**Thank you for your kind attention!**

