DDAW System test methodology for validation by the manufacturer

Clare Anderson, PhD.
Professor of Sleep and Circadian Science
Centre for Human Brain Health, School of Psychological Sciences
University of Birmingham, UK

Adjunct Professor (Research)
Monash University Accident Research Centre (MUARC), Australia

UNIVERSITY OF BIRMINGHAM

DDAW System test methodology for validation by the manufacturer

PURPOSE: Develop a standardised test procedure that will provide evidence that a DDAW system shall provide a warning to the driver at an excessive or unsafe level of drowsiness

1. How should positive and negative (drowsiness) states be induced during the validation test procedure ?

2. Should validation testing with human participants continue after the first True Positive warning is given, and if so, what requirements would determine the extended testing procedure?

3. How does the test bin change the outcome? How long should it be, and should the bins before/after the warning be considered, and why?

4. Should there be requirement for a minimum number of test runs per participant to ensure statistical significance, as opposed to being able to use a given participant once, but another participant multiple times?

UNIVERSITY OF BIRMINGHAM

Topics for Discussion

1. Induction of Positive/Negative Drowsy States for Valid Tests
2. Participant Selection and Testing Schedules
3. End of Test Decisions
4. Analysis of Data – Recommended Bin sizes

Several factors can affect test outcomes

Test Parameters
- Establishing drowsy and non-drowsy states are essential for valid trials

Trial design can affect trial outcomes
- Duration of test
- Participant selection
- Type of trial (e.g., track)

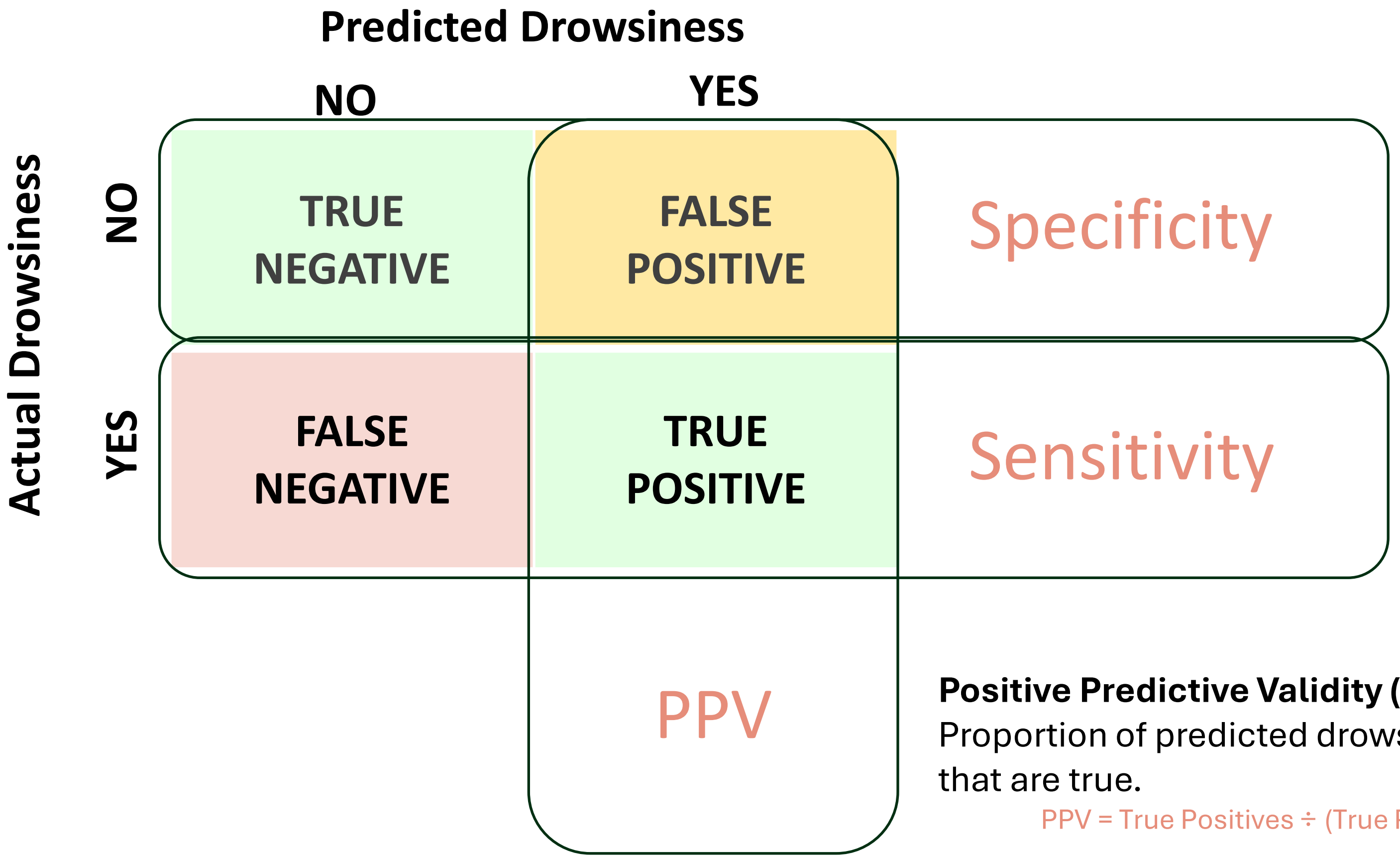
Create Drowsy/Non-Drowsy States → Determine Metrics → Conduct Validation Test → Assess Outcome

KSS8 as a drowsiness 'ground truth'
- Maps to driving outcomes and crash risk (e.g., Anderson et al., 2023)
- Strongly associated with drowsiness (long eye closures and microsleep (e.g., Manousakis et al. 2021)
- Other evidence-based measures comparable to KSS8 for detecting drowsiness

Data handling can affect trial outcome
- Data binning (window of assessment)
- Determine true positive
- Ending trial

UNIVERSITY OF BIRMINGHAM

What is needed for a valid ‘test’?

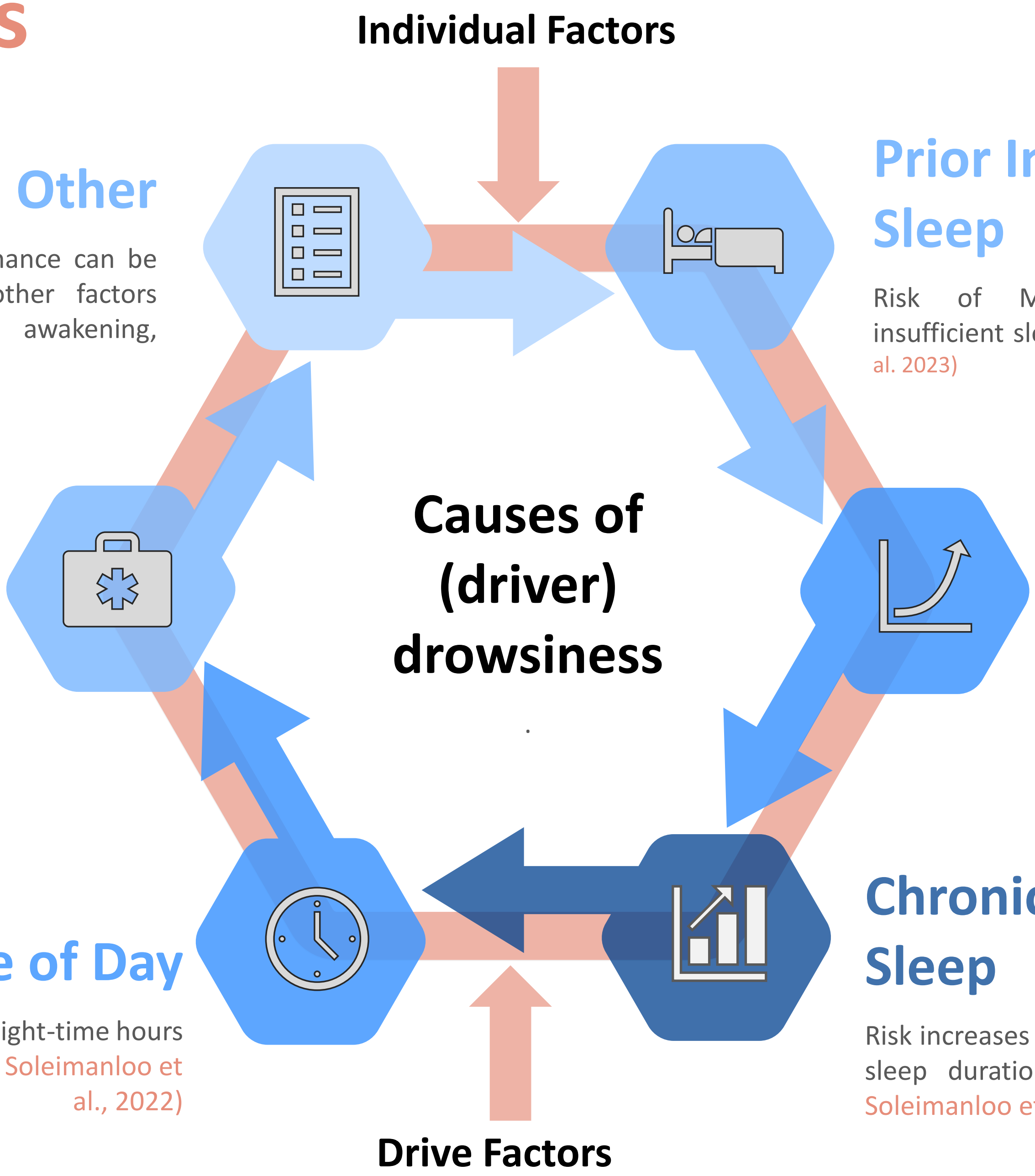


Specificity
Proportion of true negatives correctly identified
Sensitivity = True negatives ÷ (True Negatives + False Positives)

Sensitivity
Proportion of true positives correctly identified
Sensitivity = True Positives ÷ (True Positives + False Negatives)

Positive Predictive Validity (PPV
Proportion of predicted drowsiness events (e.g., Alarms) that are true.
PPV = True Positives ÷ (True Positives + False Positives)

Inducing Drowsiness



Other

Driving performance can be impaired by other factors (e.g., upon awakening, medication)

Prior Insufficient Sleep

Risk of MVC increases with insufficient sleep (<5h) (e.g. Sprajcer et al. 2023)

Sleep Disorders

Sleep disorders (esp OSA) lead to a higher risk of fall-asleep MVCs up to 8x risk.

Increased Time Awake

Risk increases with each hour of wakefulness after nine hours on shift (e.g., Shekari Soleimanloo et al., 2022)

Time of Day

Peak risk occurs during the night-time hours and afternoon (e.g., Shekari Soleimanloo et al., 2022)

Chronic Insufficient Sleep

Risk increases exponentially with each day that sleep duration is insufficient (e.g., Shekari Soleimanloo et al., 2022)

UNIVERSITY OF BIRMINGHAM

Infographic adapted from presentaitongo.com

Inducing Drowsiness (Negative/Positive States)

LIBERTY MUTUAL *(2011-12)*

Optalert®

- Shift workers of varying ages
- Day/Night shifts
- KSS, SSS
- Near-crash Events/Lane Deviations

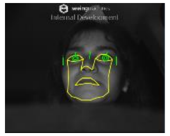
Lee et al. 2016; Anderson et al. 2023

IBAS

Liberty Mutual. INSURANCE

ARC-TRACK *(2017-19)*

Seeing Machines ™



- Younger and older drivers
- 0h versus 8h sleep
- KSS, SSS, Devices
- Near-crash Events/Lane Deviations

seeing machines

Australian Government

Australian Research Council

Cai et al. 2020, 2023; Manousakis et al. 2025

Project DRIVES (**D**ose-**R**esponse **I**n-**V**ehicle **E**valuation of **S**leepiness) *(2019-22)*

AmTech PST©

- Younger drivers
- 0h, 3h, 5h and 8h sleep
- KSS, SSS, Devices
- Near-crash Events/Lane Deviations

VICTORIA State Government — Department of Transport and Planning

TAC Transport Accident Commission

Manousakis et al. 2025

D-TECH (Drowsiness Technology) *(2021-24)*

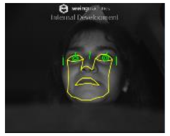
Geobox©

Cardiowheel©

PVT Workfit ©

- Middle-aged drivers
- 0h and 8h sleep
- KSS, SSS, Devices
- Near-crash Events/Lane Deviations/Seeing Machines

Seeing Machines ™



Subaru

AAA

Manousakis et al. 2025, www.aaa.com.au

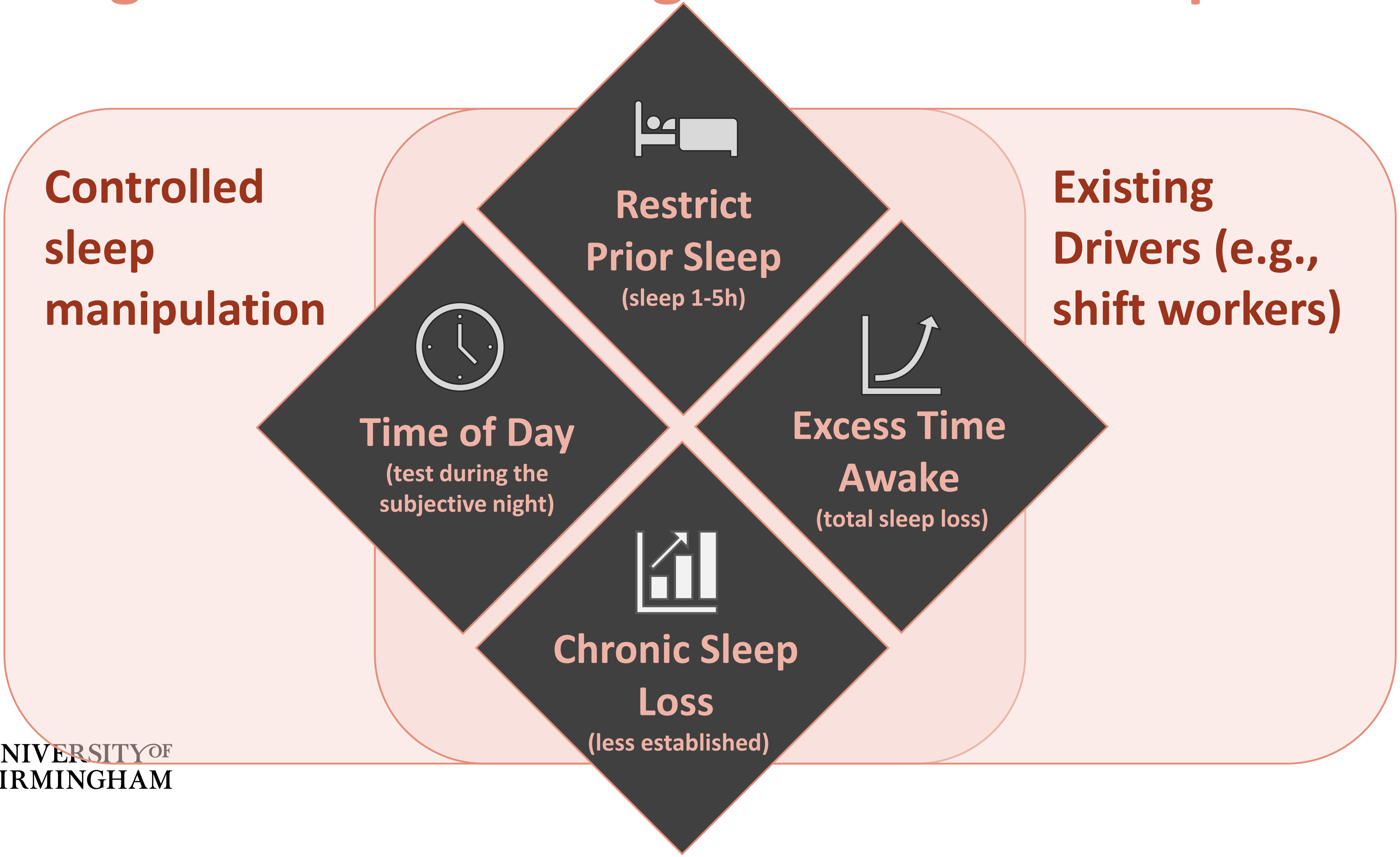
Inducing Drowsiness (Negative and Positive States) through sleep loss

Study Name	n	Age	Drive Duration (mins)	Drowsiness Induction Method	% Negative State (KSS<8)	% Positive State (KSS>8)
D-TECH	22	30-50y	120minutes	0h sleep + afternoon drive	33% pts max 56% data points	67% pts max 44% data points
ARC TRACK Y	17	21-35y	120minutes	0h sleep + afternoon drive	24% pts max 53% data points	76% pts max 47% data points
ARC TRACK O	17	50-65y	120minutes	0h sleep + afternoon drive	35% pts max 63% data points	65% pts max 37% data points
DRIVES 0h	15	21-35y	120minutes	0h sleep + afternoon drive	27% pts max 62% data points	73% pts max 38% data points
DRIVES 3h	15	21-35y	120minutes	0h sleep + afternoon drive	20% pts max 56% data points	80% pts max 44% data points
DRIVES 5h	15	21-35y	120minutes	0h sleep + afternoon drive	13% pts max 49% data points	87% pts max 51% data points

UNIVERSITY OF BIRMINGHAM

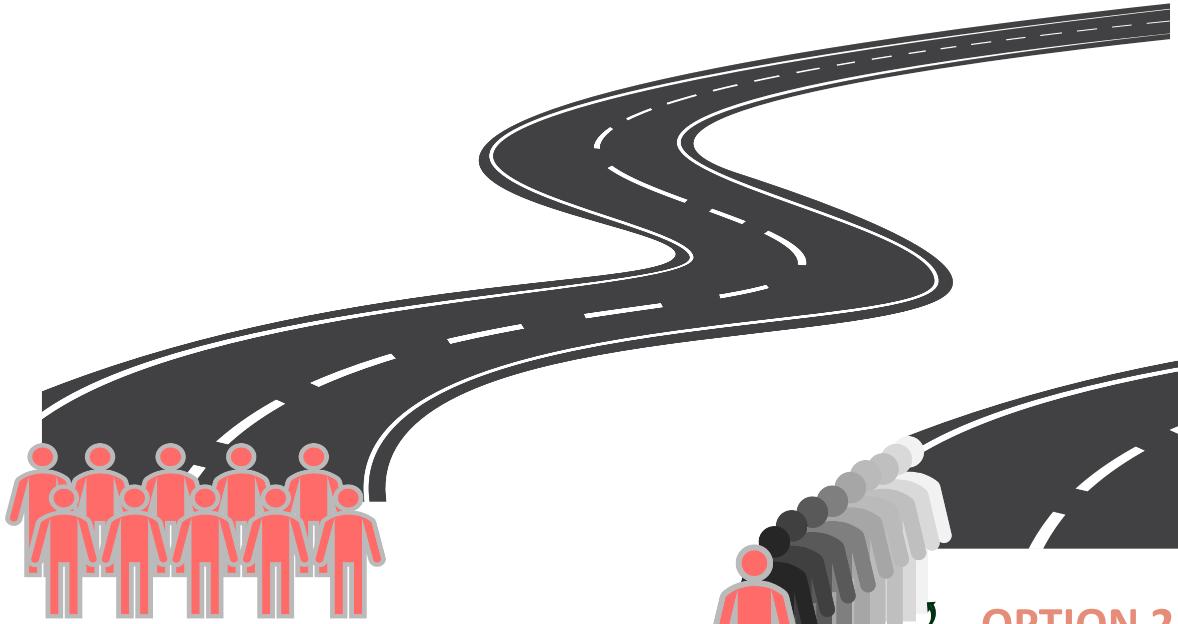
Data examples from Cai et al. 2021, Manousakis et al. 2024 and Manousakis et al. 2025

Inducing drowsiness through insufficient sleep

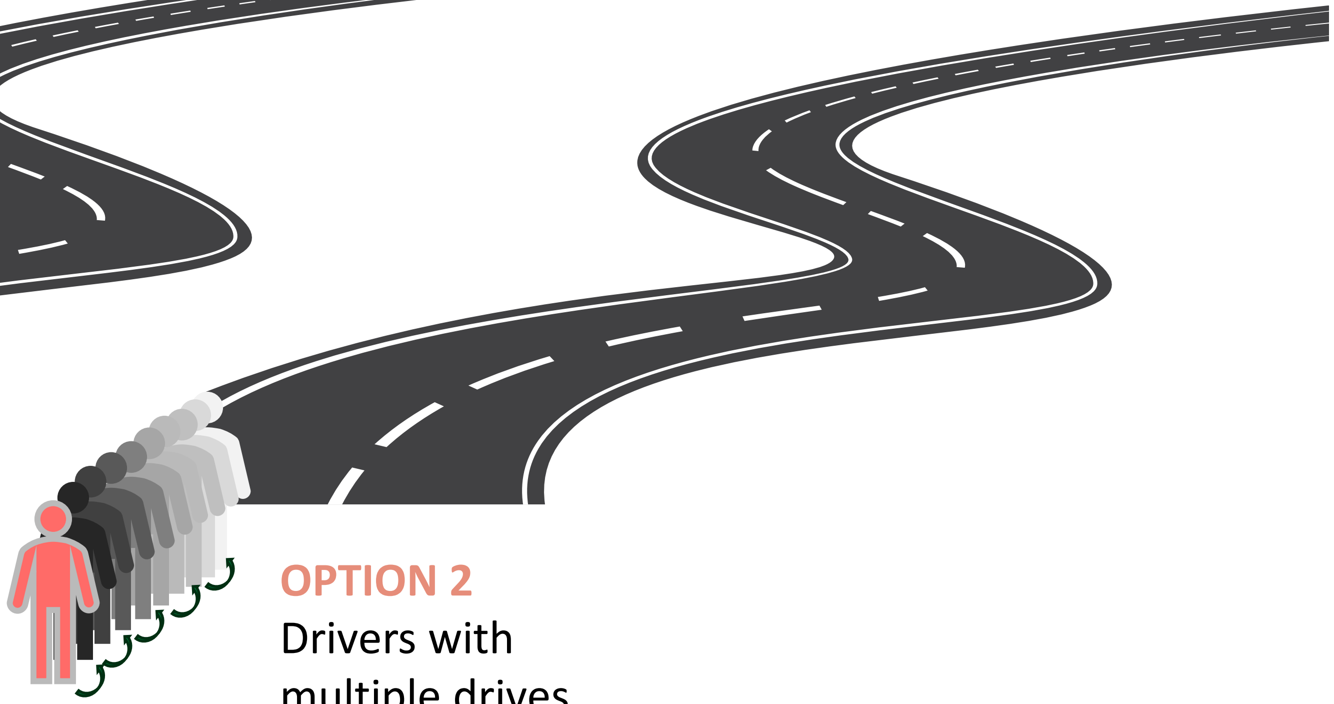


UNIVERSITY OF BIRMINGHAM

Participant Selection and Test Procedures



OPTION 1
Different Drivers



OPTION 2
Drivers with multiple drives

CRITICAL EVIDENCE ON INDIVIDUAL DIFFERENCES (Van Dongen et al., 2004)

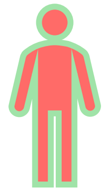
1. Large differences exist between people
2. People are highly consistent in their repeated response to sleep loss

UNIVERSITY OF BIRMINGHAM

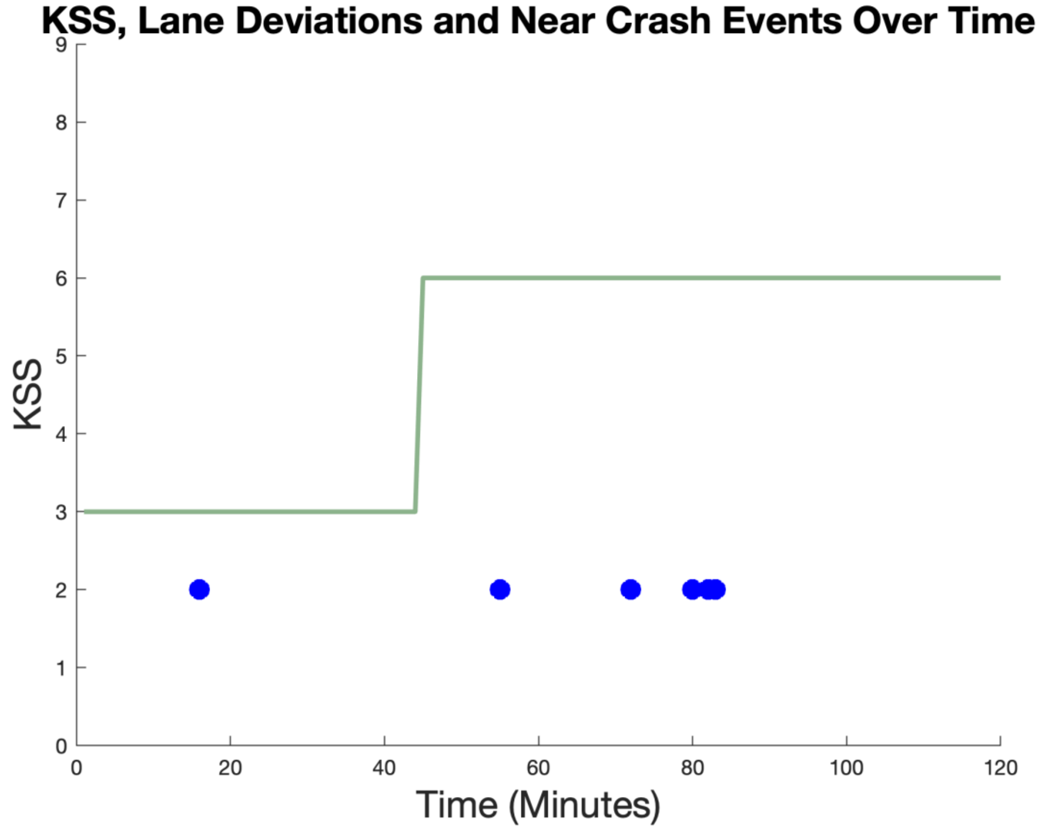
Participant Selection and Test Procedures

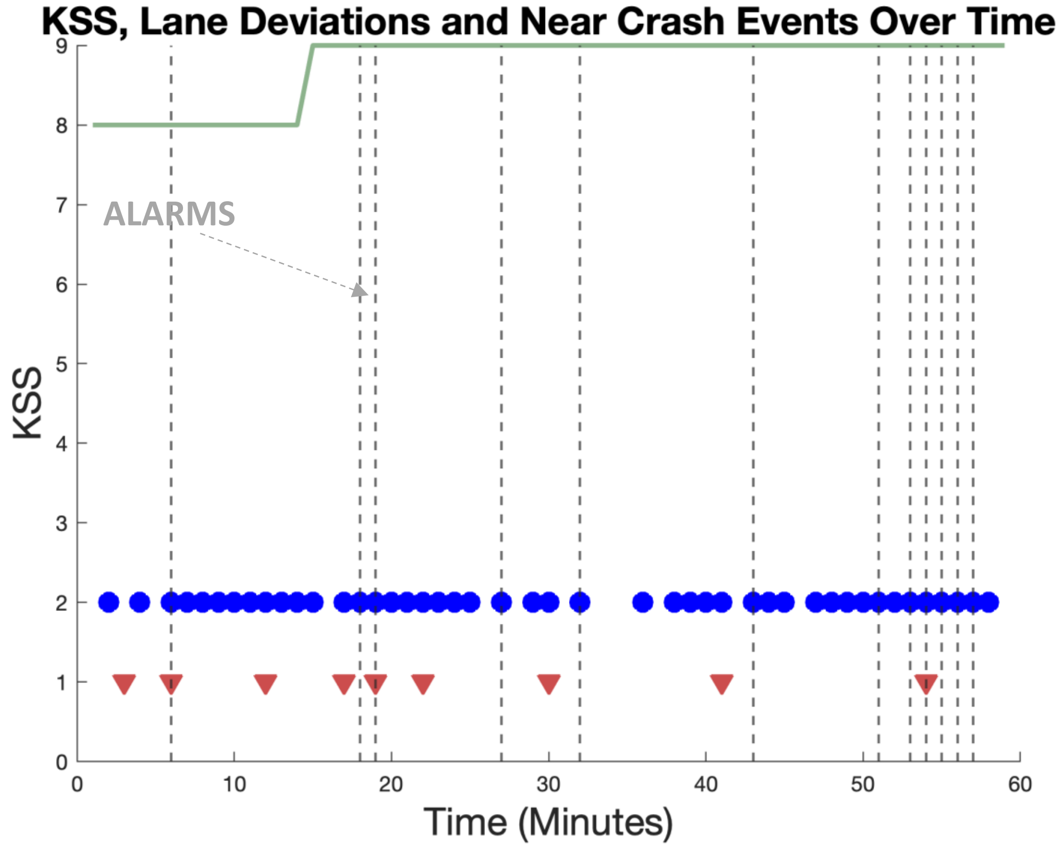
Resilient

Vulnerable



● Lane Departure
▼ Near Crash Event



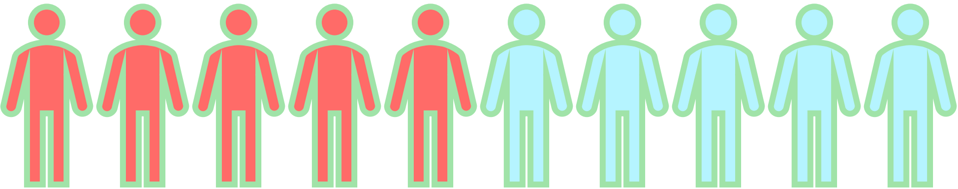


Data examples from Manousakis et al. 2024.

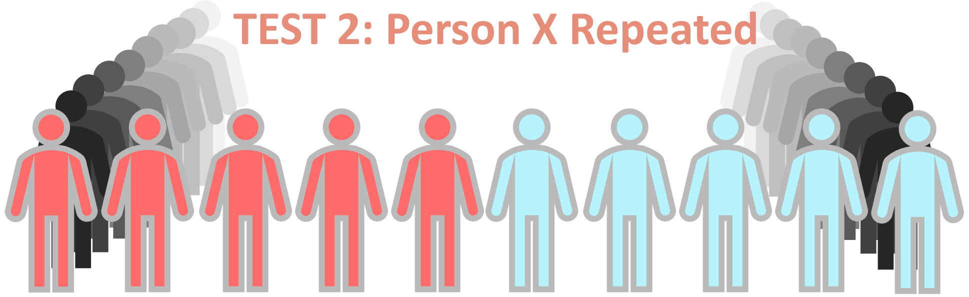
Multiple exposure skews results

Test Scenario:
Two-hour drive following a night without sleep
Classification of Drowsiness: KSS8
Device: Multiple

TEST 1: Different People*



TEST 2: Person X Repeated



KEY OUTCOMES

Repeated exposure of drivers can occur, but average sensitivity must be used.

Not averaging the data will skew of results with large error (e.g., depend on WHO is repeated and/or VALIDITY of the device)

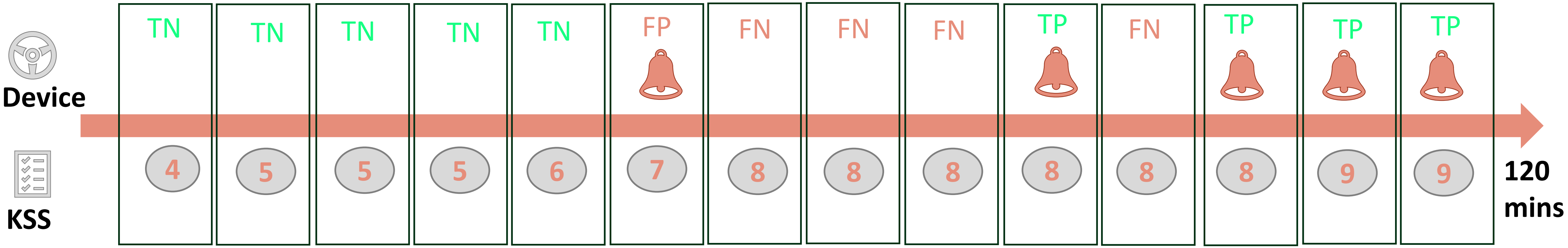
Testing sample

4.1. Each test participant shall generate at least 1 true positive or 1 false negative event as referred to in paragraphs 6.1.4. to 6.1.7. of this Appendix. **The total number, obtained by the sum of true positive events and false negative events, shall be equal to, or higher than 10.** The minimum sample size shall be 10 participants. More than one test may be run for each participant in order to acquire more data for a given participant, but the averaged data point must be used in accordance with 4.1.1

4.1.1. The sensitivity per participant shall be calculated first for each participant, then the average sensitivity and its standard deviation shall be calculated from the values of sensitivity per participant.

6.1.4.1. **Once a true positive event has occurred, all the data points after this event shall be considered irrelevant for this specific test.** If the participant restarted the test after a rest, it shall be considered a different dataset (with the same participant).

Ending the trial is an important consideration

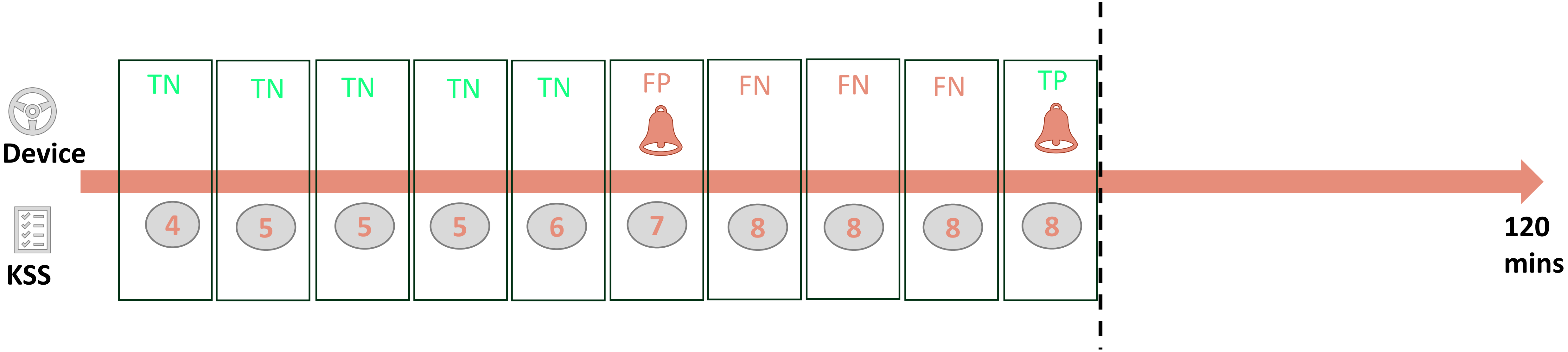


Sensitivity
Proportion of true positives correctly identified
Sensitivity = True Positives ÷ (True Positives + False Negatives)

Whole trial ➡ **4/(4+4) = 0.50 (50%)**

UNIVERSITY OF BIRMINGHAM

Ending the trial is an important consideration



Sensitivity

Proportion of true positives correctly identified

Sensitivity = True Positives ÷ (True Positives + False Negatives)

Whole trial → **4/(4+3) = 0.57 (57%)**

End TP → **1/(1+3) = 0.25 (25%)**

UNIVERSITY OF BIRMINGHAM

Ending the test: Ground Truth (KSS) vs Device

Device	Trial End	Sensitivity	Specificity	PPV
Device 1 (Lane) 	End of Drive	34%	93%	80%
Device 2 (Ocular) 	End of Drive	44%	98%	93%
Device 3 (Ocular) 	End of Drive	39%	86%	63%
Device 4 (Native) 	End of Drive	7%	99%	80%

The D-TECH Program: 2020-2024

Australian Automobile Association

Data example taken from the study described here: Manousakis, J. E. et al. (2024)

Ending the test: Ground Truth (Near Crash)

Device	Trial End	Sensitivity	Specificity	PPV
Device 1 (Lane) 	**End of Drive**	**60%**	**86%**	**40%**
	FIRST True Pos	50%	90%	14%
Device 2 (Ocular) 	**End of Drive**	**75%**	**89%**	**52%**
	FIRST True Pos	50%	92%	18%
Device 3 (Ocular) 	**End of Drive**	**68%**	**83%**	**37%**
	FIRST True Pos	25%	83%	5%
Device 4 (Native) 	**End of Drive**	**15%**	**98%**	**60%**
	FIRST True Pos	7%	98%	33%

The D-TECH Program: 2020-2024

Australian Automobile Association

Data example taken from the study described here: Manousakis, J. E. et al. (2024)

Ending the test after the first TP or later

Advantages	Disadvantages

Advantages

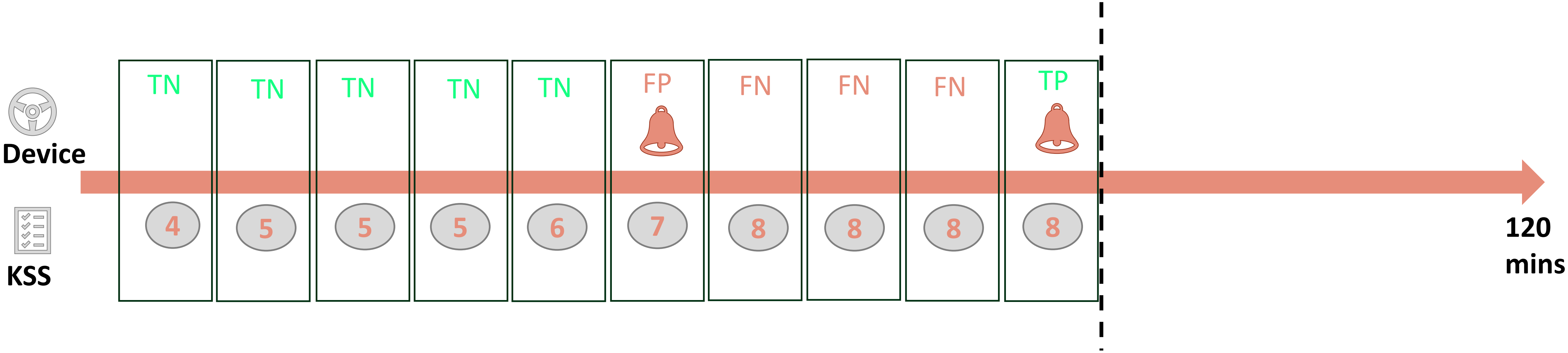
- Reduces sensitivity thus maintains high standards for safety
- Ensures standardisation for all test situations
- Reduces trial length and burden for manufacturers

Disadvantages

- Reduces sensitivity so harder for manufacturers to meet standards
- Need to monitor ground truth (KSS) and devices (alarms) in real-time
- When to end the test in the event of no TP

UNIVERSITY OF BIRMINGHAM

Ending the trial is an important consideration



QU: How easily can we satisfy the requirements?

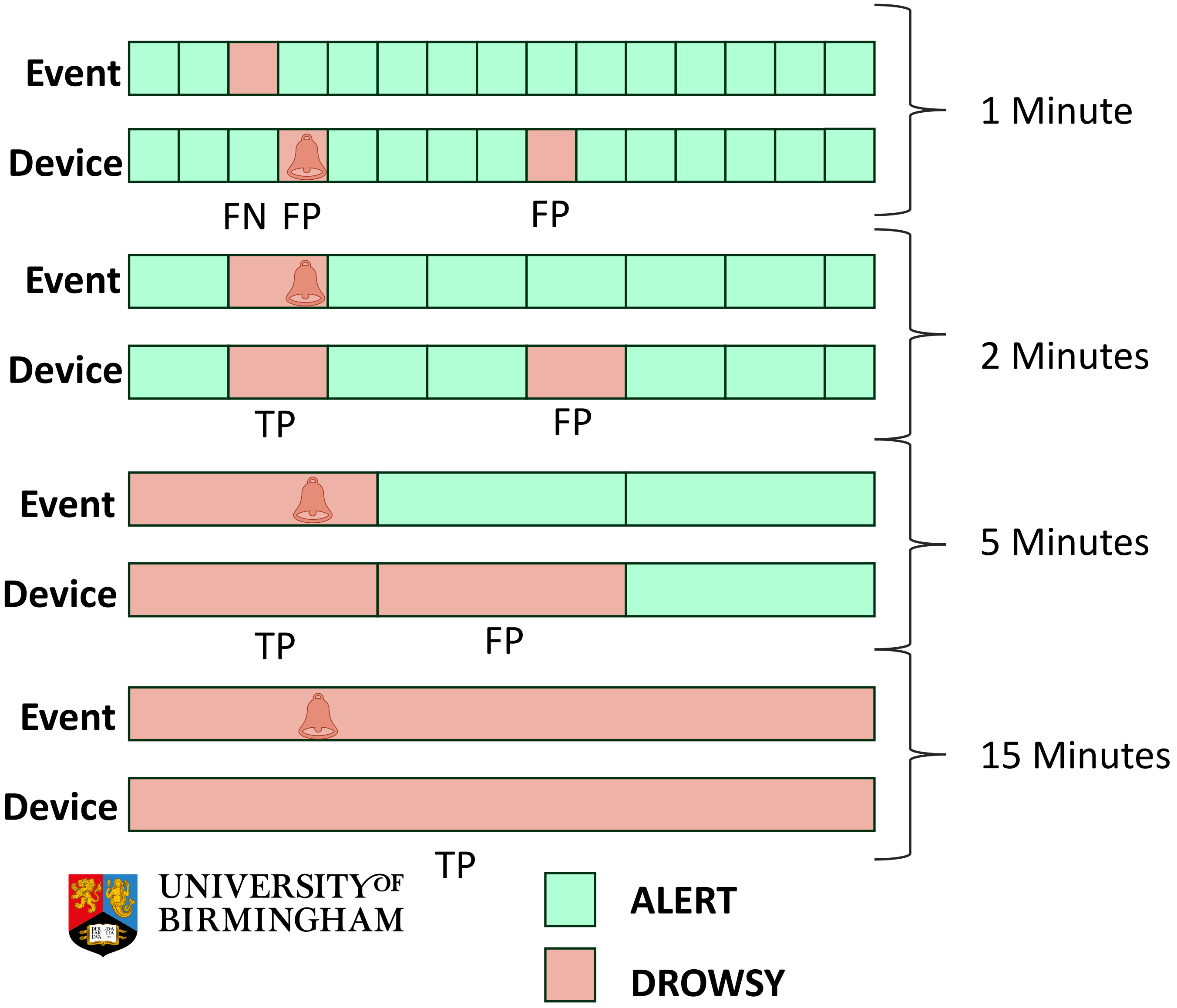
Testing sample

4.1. ..... **The total number, obtained by the sum of true positive events and false negative events, shall be equal to, or higher than 10.**

6.1.4.1. **Once a true positive event has occurred, all the data points after this event shall be considered irrelevant for this specific test.** ....

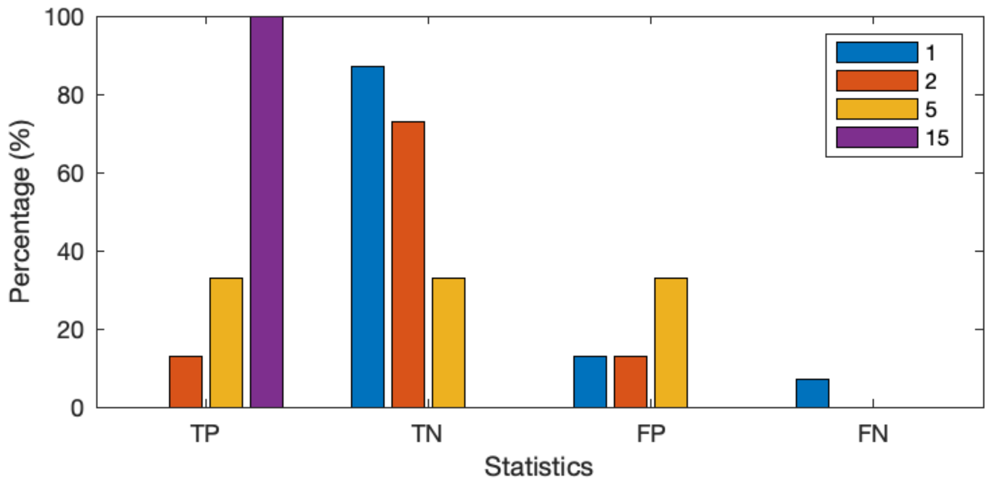
QU: Why end after the first TP?

Larger bin size increases the likelihood of a true positive

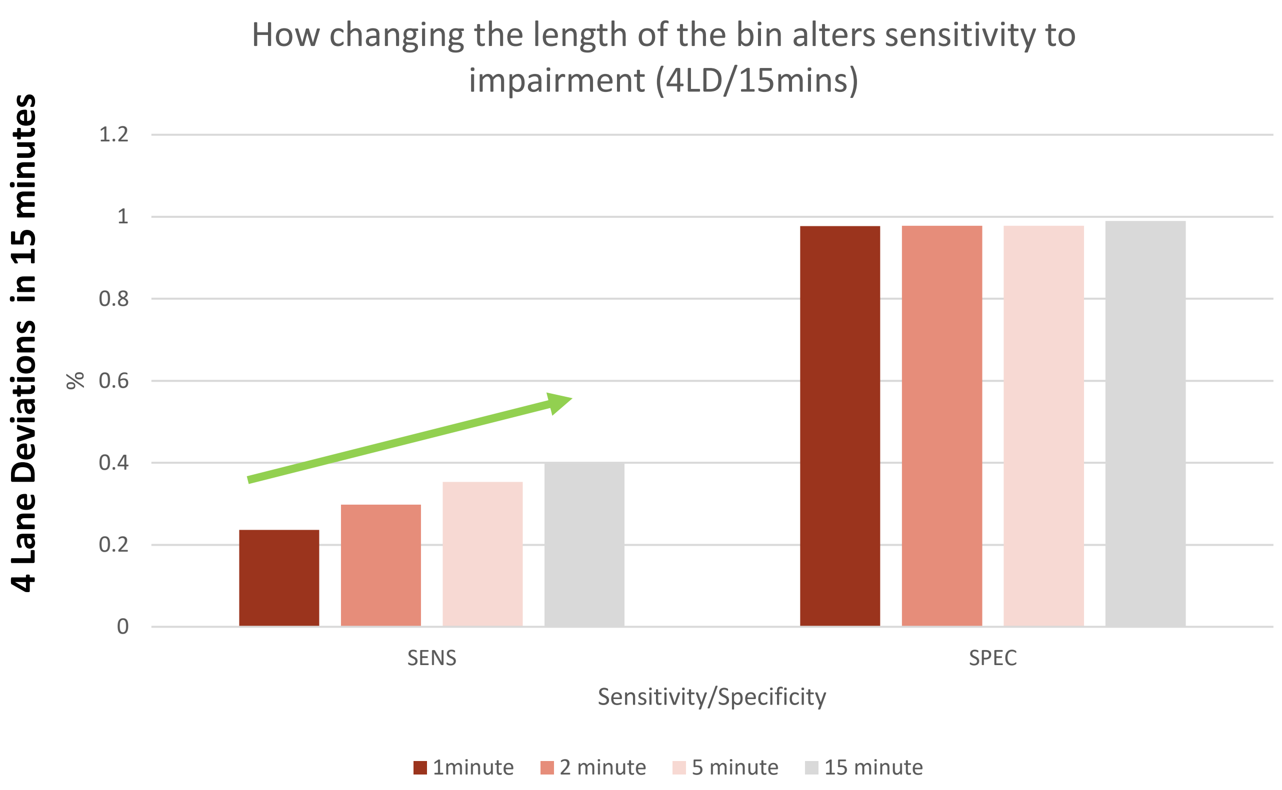


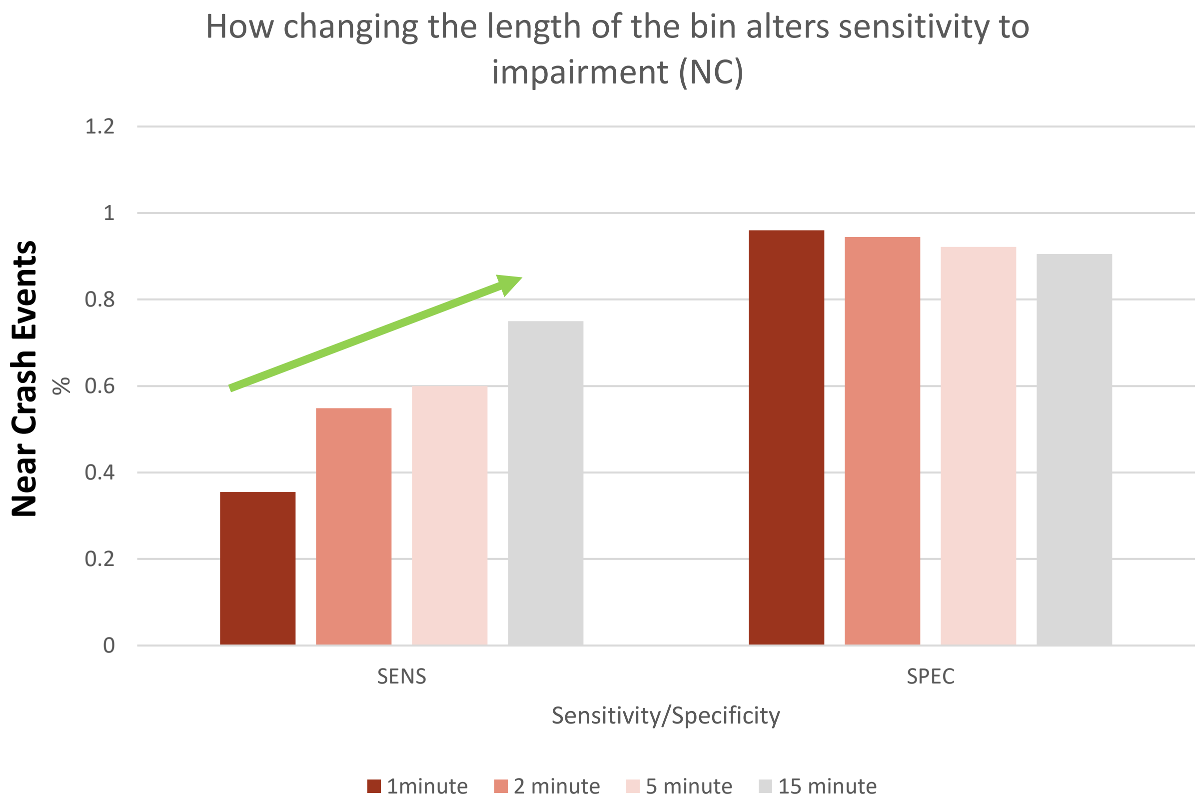
UNIVERSITY OF BIRMINGHAM

Bin Size	TP	TN	FP	FN
1	0/15	13/15	2/15	1/15
2	1/7.5	5.5/7.5	1/7.5	0/7.5
5	1/3	1/3	1/3	0/3
15	1/1	0/1	0/1	0/1



Bin size for calculations matters: a case study analysis





The D-TECH Program: 2020-2024

Australian
Automobile
Association

Data example taken from the study described here: Manousakis, J. E. et al. (2024)

Bin size for calculations matters

- Larger time windows (bins) increase the chance that the device activates in the same window (bin) as the true drowsiness event.
- Sensitivity therefore increases with larger windows (bins)
- A 15minute window (bin) can have a gap of 14 minutes between alarm and event

6.1.2. Measurements shall be obtained during the testing at regular intervals of **between approximately 5 minutes and 15 minutes**, where each measurement obtained shall be assumed to cover the previous interval.

6.1.4 Any warning from the DDAW system shall **be treated as a true positive event if the participant's previous or next rating is at a KSS of level 7 or above**. Paragraphs 6.1.6. and 6.1.7. provide further clarification on generation of true positive events.

Bin size for calculations matters

- Larger time windows (bins) increase the chance that the device activates in the same window (bin) as the true drowsiness event.
- Sensitivity therefore increases with larger windows (bins)
- A 15minute window (bin) can have a gap of 14 minutes between alarm and event: **is this (and longer) acceptable?**
- A smaller temporal resolution of the time window allows for a system to ‘warn’ the driver in a timely manner (Watling et al., 2021)
- Standardization (or capping) of the bin size is recommended, and an evaluation of existing data to systematically determine the influence of these ‘decisions’

Several factors can affect test outcomes

Test Parameters
- Establishing drowsy and non-drowsy states are essential for valid trials

Trial design can affect trial outcomes
- Duration of test
- Participant selection
- Type of trial (e.g., track)

Create Drowsy/Non-Drowsy States → Determine Metrics → Conduct Validation Test → Assess Outcome

KSS8 as a drowsiness ‘ground truth’
- Maps to driving outcomes and crash risk (e.g., Anderson et al., 2023)
- Strongly associated with drowsiness (long eye closures and microsleep (e.g., Manousakis et al. 2021)
- Other evidence-based measures comparable to KSS8 for detecting drowsiness

Data handling can affect trial outcome
- Data binning (window of assessment)
- Determine true positive
- Ending trial

UNIVERSITY OF BIRMINGHAM

Additional Evidence is Needed

Summary

1. Induction of drowsy states can be done in many ways and largely relate to prior sleep.
2. Testing a variety of drivers is recommended, and repeating drivers should only occur if averages are used.
3. Ending after the first TP can influence values but generally reduces sensitivity (doesn't compromise safety).
4. Additional systematic checks are required on the operational outcomes of the analyses.
5. Time window for assessment (e.g., bin size) of classification (e.g., TP, FN, etc) is important and requires more discussion/evidence.

DDAW System test methodology for validation by the manufacturer

PURPOSE: Develop a standardised test procedure that will provide evidence that a DDAW system shall provide a warning to the driver at an excessive or unsafe level of drowsiness

1. How should positive and negative (fatigue) states be induced during the test procedure for validation?

   Evidence suggests many states can be used to ensure **both** states are observed

2. Should there be requirement for a minimum number of test runs per participant to ensure statistical significance, as opposed to being able to use a given participant once, but another participant multiple times?

   Repeated participants can introduce error (through averaging), but essential this occurs.

3. Should validation testing with human participants continue after the first True Positive warning is given, and if so, what requirements would determine the extended testing procedure?

   Advantages exist for ending the test after the first TP, but sensitivity will be reduced

4. How does the test bin change the outcome? How long should it be, and should the bins before/after the warning be considered, and why?

   Wider bin sizes increase the likelihood of a TP. Consider restricting/capping the bin size to manage safety implications

UNIVERSITY OF BIRMINGHAM

Acknowledgements

COLLABORATORS AND RESEARCH SUPPORT

Project Leaders/Post Doctoral Fellows

Sophie Mason, Ph.D.*
Jessica Manousakis, Ph.D.*
Suzanne Ftouni, Ph.D.
Katy Jeppe, Ph.D.
Anna Cai, Ph.D.
Charmaine Diep, Ph.D.
Madelaine Sprajcer, Ph.D.
Michael Lee, Ph.D.
Jennifer Cori, Ph.D.

Collaborators

Mark Howard, M.D., Ph.D.
Jim Horne, Ph.D.
Sally Fergusson, Ph.D.
Anjam Naweed, Ph.D.
Charles Czeisler, Ph.D.
Bill Horrey, Ph.D.
Mike Lenne, Ph.D.
Shantha Rajaratnam, Ph.D.

All students and research support staff in the Division of Sleep Medicine (Brigham and Women's Hospital), the Monash Sleep and Circadian Laboratory, and the METEC Driver Training Facility.

All the participants who make the work possible.

* Running data processing/analyses for workshop

FUNDING SUPPORT

VICTORIA State Government — Department of Transport and Planning

TAC Transport Accident Commission

Australian Government — Australian Research Council

seeing machines

CRC Alertness Safety and Productivity

Australian Government — Department of Industry, Innovation and Science — Business Cooperative Research Centres Program — business.gov.au 13 28 46

Australian Automobile Association

OPTALERT®

UNIVERSITY OF BIRMINGHAM

References

Anderson, C. et al. Feeling sleepy? stop driving-awareness of fall asleep crashes. Sleep 46 (2023).

Cai, A.T. et al. (2021). On-road driving impairment following sleep deprivation differs according to age. Scientific Reports. 11: 21561.

Lee, M. L. et al. High risk of near-crash driving events following night-shift work. Proc Natl Acad Sci U S A 113, 176-181, (2016).

Manousakis, J. E. et al. From Drift to Danger: Predicting Near Crashes from Lane Deviations in Drowsy Drivers. Under Review (Preprint SSRN) (2025).

Manousakis, J. E. et al. Evaluation, Validation and Comparison of Fatigued Driving Monitoring Systems Final Report. (https://www.aaa.asn.au/library/evaluation-validation-and-comparison-of-fatigued-driving-monitoring-technologies/, 2024).

Manousakis, J. E., Mann, N., Jeppe, K. J. & Anderson, C. Awareness of sleepiness: Temporal dynamics of subjective and objective sleepiness. Psychophysiology 58, e13839, (2021).

Sprajcer et al. (2023). How Tired is Too Tired to Drive? A Systematic Review Assessing the Use of Prior Sleep Duration to Detect Driving Impairment. Nat Sci Sleep. 15:175-206.

Shekari Soleimanloo et al. (2022). The association of schedule characteristics of heavy vehicle drivers with continuous eye-blink parameters of drowsiness. Transportation Research Part F: Traffic Psychology and Behaviour. 90: 485-499.

Watling, Hasan & Larue (2021). Sensitivity and specificity of the driver sleepiness detection methods using physiological signals: A systematic review. Accident Analysis & Prevention, Volume 150, 105900.

DDAW System test methodology for validation by the manufacturer

Clare Anderson, PhD.
Professor of Sleep and Circadian Science
Centre for Human Brain Health, School of Psychological Sciences
University of Birmingham, UK

Adjunct Professor (Research)
Monash University Accident Research Centre (MUARC), Australia

UNIVERSITY OF BIRMINGHAM