Towards better subjective sleepiness ground truth data quality

Raimondas Zemblys[1], Christer Ahlström[2,3], Anna Anund[2,4], Svitlana Finér[1]

[1]Smart Eye, AB, Gothenburg, Sweden; [2]Swedish National Road and Transport Research Institute, Linköping, Sweden; [3]Department of Biomedical Engineering, Linköping University, Linköping, Sweden; [4]Rehabilitation Medicine, Linköping University, Linköping, Sweden

Introduction

The Karolinska Sleepiness Scale (KSS)

- KSS is a subjective scale that has been found to correlate with objective and behavioral measures of sleepiness.
  - however, the subjective feeling does not always reflect the objective sleepiness level, and the anchored scale might be interpreted differently by different drivers.
- When using KSS to develop and evaluate sleepiness detection systems, ***accurate*** and ***absolute*** ratings are essential
  - therefore, all drivers must be trained to have the same understanding of the scale.

Objectives

- Investigate how different KSS training protocols affect the ratings.
- Evaluate whether the drivers can evaluate themselves consistently multiple times.
- Provide recommendations for the best practices in collecting KSS ground truth data.

Datasets

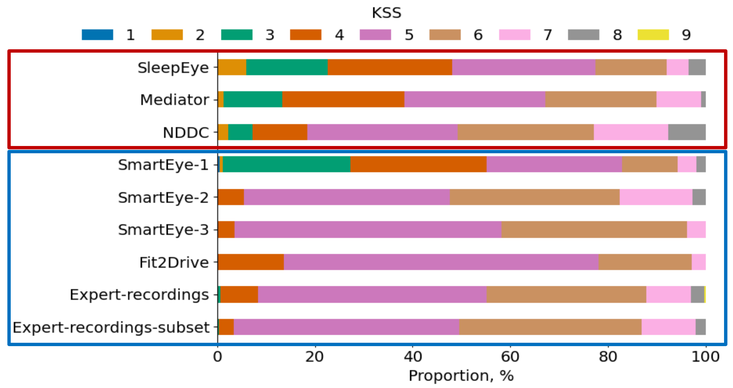
- Datasets were generated by a series of independent studies.
- In all datasets KSS was retrospectively self-reported every 5th minute while driving.
- All except *Expert-recordings* datasets were recorded in a controlled field setting:
  - one trip per driver, daytime, predefined, mostly highway route with a test leader present in the vehicle.
- The *Expert-recordings* dataset was naturalistic driving data that included multiple drives per driver (up to 120 hours)
  - *Expert-recordings-subset* consists of one (the one closest to 3 hours) drive per driver.

Method

- Datasets were divided into two categories:
  - **"KSS-only instruction" datasets.** In *SleepEye*, *Mediator*, and *NDDC* datasets drivers received training on the KSS scale, including the anchors.
  - **"Additional KSS instruction" datasets**. In the rest, KSS labels were amended with explanations and examples on how to interpret and report KSS.

Results

Understanding of KSS

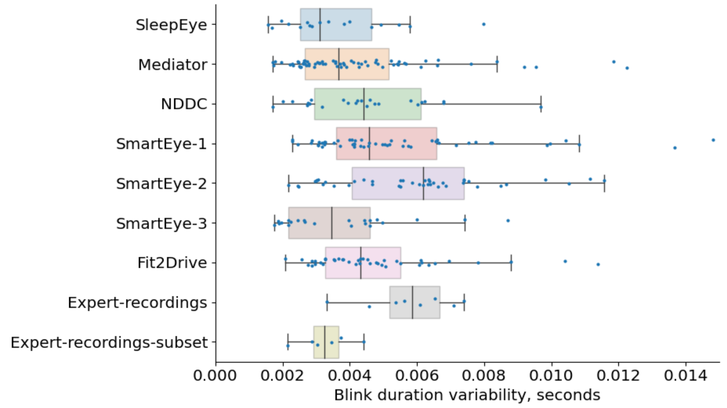


- Most of the annotations were KSS 5–6 in the "Additional KSS instruction" datasets, while in the "KSS-only instruction" datasets distributions were wider, including more KSS 3–4 and KSS 7–8 ratings
  - except for the *SmartEye-1* - most likely the consequence of KSS 3 being further described as "the peak of your day and you feel as alert and awake as you ever do".
- The "Additional KSS instruction" datasets had fewer KSS level changes compared to the “KSS-only instruction” datasets
  - most of the changes were ±1 KSS level in all datasets.
- Reporting the descriptive KSS label (as in *Fit2Drive*) likely leads to a more accurate ratings than merely reporting the number
  - *SmartEye-2* and *SmartEye-3* were recorded in US using same instructions and had many KSS values that indicated sleepiness (KSS 6–7). In *SmartEye-3* drives were approx. 1 hour therefore we did not expect drivers to become sleepy.
  - the *Fit2Drive* recordings had fewer KSS 6-7 ratings despite being recorded in Sweden and lasting approx. 3 hours.

Results

Measure stability

- Operationalized by stability evaluation of blink durations:
  - measures whether drivers were able to evaluate themselves consistently multiple times;
  - estimated by first calculating the robust mean of blink durations within 5-minute sliding windows. Then all means were aggregated within each KSS level, and the mean of interquartile ranges was calculated;
  - the lower the metric, the more consistent the driver’s blink durations are with respect to the subjective KSS ratings.



- Blink duration stability improved when the participants were forced to practice and learn the KSS by heart:
  - by using a sleep diary (*SleepEye*, *Mediator*), or
  - when description-based KSS annotations were made (*Fit2Drive*, part of *Expert recordings* datasets).
- Metric is sensitive to the recording duration:
  - short drives might appear more stable (*SmartEye-2* vs *SmartEye-3*, *Expert-recordings* vs *Expert-recordings-subset*).

Conclusions

- KSS training instructions have huge impact on the reported KSS levels:
  - similar training instructions resulted in similar KSS distributions across datasets and less KSS level changes throughout the drive(s);
  - not learning the scale well led to developing an own understanding of the scale, which resulted to confusion and less accurate ratings.
- Description-based ratings reflect driving conditions best and align state understanding across the drivers better:
  - such training resulted in reasonable KSS distributions and dynamics, while also giving stable blink durations over different KSS levels.
- Rigorous learning of the scale, in combination with a standardized data collection protocol, is needed to get replicable results;
  - however, more controlled studies are needed to remove the many confounds we had in and between our datasets (different routes, driving conditions, driver backgrounds, etc.).