Eur. Transp. Res. Rev. (2015) 7: 38
DOI 10.1007/s12544-015-0188-y

ORIGINAL PAPER

Video-based observer rated sleepiness versus self-reported subjective sleepiness in real road driving

Christer Ahlstrom[1] · **Carina Fors**[1] · **Anna Anund**[1] · **David Hallvig**[2]

Received: 9 September 2014 / Accepted: 16 November 2015 / Published online: 23 November 2015

Abstract
Purpose Observer-rated sleepiness (ORS) based on video recordings of the driver's face is often used when analysing naturalistic driving data. The aim of this study is to investigate if ORS ratings agree with subjective self-reported sleepiness (SRS).
Methods Forty raters assessed 54 video-clips showing drivers with varying levels of sleepiness. The video-clips were recorded during a field experiment focusing on driver sleepiness using the same cameras that are typically used in large-scale field studies. The weak results prompted a second test. Ten human factors researchers made pairwise comparisons of videos showing the same four participants in an alert versus a very sleepy condition. The task was simply to select the video-clip where the driver was sleepy.
Results The overall average percentage of video segments where ORS and SRS matched was 41 % in Test 1. ORS 0 (alert) and ORS 2 (very sleepy) were easier to score than ORS 1 and it was slightly harder to rate night-time drives. Inter-rater agreement was low, with average Pearson's r correlations of 0.19 and Krippendorff's alpha of 0.15. In Test 2, the average Pearson's r correlations was 0.35 and Krippendorff's alpha was 0.62. The correspondence between ORS and SRS showed an agreement of 35 %.
Conclusions The results indicate that ORS ratings based on real road video recordings correspond poorly with SRS and have low inter-rater agreement. Further research is necessary in order to further evaluate the usefulness of ORS as a measure of sleepiness.

Keywords Observer rated sleepiness · Video annotation · Driver sleepiness · Field study

✉ Christer Ahlstrom
christer.ahlstrom@vti.se

[1] Swedish National Road and Transport Research Institute (VTI), S-58195 Linköping, Sweden

[2] Stress Research Institute, Stockholm University, Stockholm, Sweden

1 Introduction

A recent trend in studies of driver sleepiness is to carry out large-scale naturalistic data collections [1–3]. In a naturalistic driving study, a large number of volunteer participants drive a vehicle (usually their own) for an extended period of time (usually 6 to 12 months, and sometimes even longer) during their normal everyday activities. The vehicles are instrumented with unobtrusive data acquisition systems that continuously record the behaviour of the vehicle (e.g. speed and lane position), the behaviour of surrounding road users, and the drivers' behaviour (e.g. where they are looking). The advantage of these naturalistic driving studies, from a driver sleepiness perspective, is the possibility to study the extent to which sleepiness contributes to safety critical incidents [1, 4].

Sleepiness is the transitional state between wakefulness and sleep, a state that is easily perceived but difficult to measure [5, 6]. Several approaches to measure the effects of driver sleepiness have been explored in the literature. These include physiological recordings and their scoring [e.g. 7, 8], blinking activity and eye movements measured by cameras [e.g. 9], measures of driving performance [e.g. 10–12], self-reported sleepiness, SRS [e.g. 7, 11, 13, 14], and observer-rated sleepiness, ORS [e.g. 15–17].

In controlled simulator studies and field studies on real roads, subjective self-reported sleepiness ratings have been found to be useful [18]. The technique is cheap and easy to use, and it is unaffected by the noisy environment found in a

vehicle. Above all, subjective sleepiness ratings are the driver sleepiness measure that is least affected by between individual variations, also in comparison with physiological sleepiness indicators [13]. In naturalistic driving studies, sleepiness ratings based on self-reports or on physiological data are too intrusive. Instead, unobtrusive techniques such as ORS and camera-based solutions are used. ORS is today the prevailing method of choice.

The underlying idea of ORS is simply to let an observer estimate the level of sleepiness experienced by the driver based on behavioural signs of sleepiness such as the driver's facial expression, body movements, postural changes and duration of eyelid closures [19]. ORS estimations may be carried out in two different ways: (i) real-time ORS performed by test leaders accompanying the driver in the car [15], and (ii) post-hoc ORS based on video recordings of the drivers' face [20–22]. The second approach is used in naturalistic driving studies since no test leader is present in the vehicle.

In laboratory experiments, untrained observers have been able to accurately distinguish between photographs of individuals who had been sleep deprived or had a normal night of sleep [23]. It has also been found that observers perceived sleep deprived individuals as having more hanging eyelids, redder eyes, more swollen eyes, darker circles under the eyes, paler skin, more wrinkles/fine lines, and more droopy corners of the mouth [24]. These findings suggest that humans are sensitive to facial cues and provide support that it is possible for an observer to instantaneously rate driver sleepiness. In a driving context, ORS was first described by Wierwille and Ellsworth [17] and was later refined by Wiegand et al. [16]. Here trained observers evaluated drivers' sleepiness level using video recordings of the drivers' face. The results showed good intra-rater and inter-rater reliability, however, the raters' estimations were only compared with the average rated estimation for each clip [17], or in the refined version, with an experienced rater's ratings [16]. Consequently, these studies did not indicate the extent to which ORS reflects the sleepiness level experienced by the driver. Using the intrinsic measure of self-rated sleepiness (Karolinska Sleepiness Scale, KSS) as a comparison, it has been shown that real-time ORS reflect the general variations of self-rated sleepiness in drivers on a global level, but also that ORS was not consistently sensitive to abrupt changes in driver sleepiness at the 5-min level [15].

The overall aim of this paper is to shed some light on the issue whether an external observer is able to assess sleepiness within a real life context based on video recordings of the drivers face. More specifically, the correspondence between observer rated sleepiness and self-reported sleepiness, and the ability of an observer to recognize a sleepy driver, will be investigated.

The paper is structured as follows. In the first section, the different measures of sleepiness that are used in this paper are outlined. Subsequently, the methodology, results and discussion of two separate analyses based on data from two previously acquired data sets are presented. Test 1 compares post-hoc ORS with SRS in a real road driving setting. Since the outcome of Test 1 was inconsistent with the inter-rater results presented by Wierwille and Ellsworth [17] and Wiegand et al. [16], a second test was conducted. Test 2 simplified the rating task with the aim to see if it is at all possible to rate sleepiness based on video recordings. Paired videos showing the same driver in two conditions, severely sleepy versus alert, were rated. The task was simply to select the video in the pair showing the sleepy driver. The final section of the paper discusses the implications of the two experiments.

2 Sleepiness rating scales

Two different rating scales were compared in this study, three level SRS (based on KSS, see Section 2.1) and three level post-hoc ORS (see Section 2.2). The subjective ratings were also supported by two other measures of sleepiness, blink duration in Test 1 and 2, and the psychomotor vigilance test (PVT) in Test 2. Unfortunately, PVT was not included in Test 1. The reason for backing up SRS with other measures of sleepiness was that self-reported sleepiness suffers from methodological limitations. For example, some individuals may deny being sleepy or may lack the ability to accurately evaluate their biological cues of sleepiness. The evaluation of ORS versus SRS should therefore be seen as a comparison rather than as a classification where one is truer than the other. That being said, KSS is the measure of driver sleepiness that is least affected by between individual variations, also in comparison with physiological sleepiness indicators [13]. Under the present circumstances, KSS is likely to be the most reliable and sensitive approach available [18].

Increased levels of sleepiness are associated with longer blink durations [9, 11], longer mean reaction times in PVT, and higher percentages of lapses/misses in PVT [25–27]. Here, blink durations were calculated as the duration between half the amplitude of the blink onset to half the amplitude of the blink offset in a recorded electrooculogram (EOG) according to Jammes et al. [28]. The PVT was set up according to Loh et al. [29], with random stimuli onsets with an interval of 2–10 s between stimuli, a maximum stimulus duration of 2 s, and a total test duration of 10 min.

2.1 Self-reported sleepiness scale

KSS was used to capture the drivers' experience of sleepiness. The scale has nine levels: 1 – extremely alert, 2 – very alert, 3 – alert, 4 – rather alert, 5 – neither alert nor sleepy, 6 – some signs of sleepiness, 7 – sleepy, no effort to stay awake, 8 – sleepy, some effort to stay awake, and 9 – very sleepy, great

effort to keep awake, fighting sleep [30]. The nine KSS levels were transformed to a three level scale according to Ahlstrom et al. [31] to match the three ORS levels. KSS 1–5 form SRS 0, KSS 6–7 form SRS 1 and KSS 8–9 combine to form SRS 2.

2.2 Observer rating scale

A three-graded scale was used for ORS (see Table 1). This ORS scale has been in development over the last few years by researchers at VTI and the Stress Research Institute at Stockholm University [15], based on the work by Wierwille and Ellsworth [17]. The objective of the scale was to describe behaviours that characterize sleepy driving, inspired by the study of Rogé et al. [19]. The observed behaviours are classified into three categories: eye-related behaviours (e.g. long eye closure and slow blink rate), facial movements (e.g. yawning) and body movements (e.g. stretching and moving trunk forwards and backwards). The three levels represent 0 = Alert; 1 = Sleepy, no effort to stay awake, and 2 = Very sleepy, great effort to keep awake. The original scale developed by Wierwille and Ellsworth [17] used five categories, however, in order to reduce error variance it was decided to decrease the number of categories to three.

Table 1 Description of the ORS

ORS 0: Alert
Blink: normal
Yawn: no
Body position: sitting still
Body movements: normal
ORS 1: Sleepy, no effort to stay awake
Blink: sporadic periods of long eyelid closure (followed by increased level of blink frequency)
Yawn: some
Body position: some situations with changing position – e.g. stretching
Body movements: some – arms, legs, scratching, rubbing eyes
ORS 2: Very sleepy, great effort to keep awake
Blink: half-closed eyes, empty gaze
Yawn: yes
Body position: yes - change often, stretch, slumped, hanging
Body movements: yes - e.g. head nodding

3 Test 1

3.1 Methods

A set of 54 1-min video clips were retained from an earlier experiment which is described in detail by Fors et al. [32]. The videos were recorded from 20 drivers, selected randomly from the register of vehicle owners, participating in a field experiment, both in an alert condition and in a sleep deprived condition. They drove on a motorway for approximately 160 km in an instrumented vehicle equipped with CAN loggers, GPS, several video cameras and eye tracking. The data logger and the sensors were identical to those used in the Swedish Vehicle Management Centre in euroFOT (http://www.eurofot-ip.eu/). The video clips are representative of the video quality in the euroFOT data. Unfortunately, this means that the video quality is sometimes low, especially at nighttime. Figure 1 shows one sample from nighttime driving (left) and one from daytime driving (right). The over exposure in the frame from nighttime driving and the slight under exposure in the daytime frame is a result of a compromise between the two lighting conditions. The resolution of the videos were 352 × 288 pixels.

For safety reasons the vehicle was equipped with dual command, and a co-driver accompanied each drive. This makes the current setup different from field operational tests and naturalistic driving since a test leader was present in the car. In addition, recordings of physiology (EEG, EOG and ECG) and KSS (in 5 min intervals) were acquired. The results from the EEG and ECG are not used in the current paper. The study was approved by the ethical committee at Linköping University.

Forty voluntary participants (aged 39 ± 18 years, 20 women) were recruited to take part in the study as raters. At arrival, the (untrained) raters were given a short introduction to the upcoming task and given a written description of the method and the procedure. Then, they were shown a sequence of 10 video clips which they were instructed to rate according to the strategies outlined in Table 1. After each clip the SRS value was shown. This was to provide an anchor to help the raters to calibrate their ORS ratings. The raters were instructed to choose the most fitting ORS level, meaning that all criteria did not have to be fulfilled in order to choose a certain level. They were also instructed that the different indicators (blink, yawn, body position and body movements) should be seen as prototypical clues of what to look for in the video. Next the raters were allowed to ask questions. Following that, the actual trial commenced. The video clips were shown on a projection screen in a dark room to 20 participants at a time, and the participants wrote down their ratings individually. Obviously, the SRS values were not revealed in the actual trial.

In between each clip there was a 10 s pause where the raters were reminded to estimate the sleepiness level and also to rate the confidence in their ORS estimations. The level of confidence was given on a scale from 1 to 10, where 1 is “no idea” and 10 is “absolutely certain”. In between the three sessions of 18 clips, the raters took a break for 10 min and were served refreshments.

All daytime clips were recorded in daylight whereas all night-time clips were recorded in darkness. The selected video

Fig. 1 Examples of the video films; night-time driving (*left*) and daytime driving (*right*)

clips were distributed evenly across day/night, and low/medium/high SRS. Video clips corresponding to SRS 0 – 2 were characterised by increasing blink durations (mean ± sd) of 104 ± 22, 119 ± 24 and 126 ± 30 ms respectively. The video clips were balanced on drivers to the best possible extent, constrained by the fact that only some drivers estimated a more severe level of sleepiness. The raters were informed that the video clips had been randomly selected and that there were sleepy drivers in the daylight videos and alert drivers in the darkness videos. In order to control for learning effects of the raters, the segments were placed in three different sequences such that each sequence contained an equal distribution of video segments for each SRS level and for day compared to night. For 20 of the 40 raters, the three sequences were shown in reversed order.

3.2 Statistical analyses

The analyses involve descriptive statistics such as accuracy, reliability and confusion matrices to get a general picture of how well ORS and SRS match. If ORS and SRS matches, the confusion matrix will have large values in the diagonal, while mismatches will be revealed in the off-diagonal entries. This information is further quantified as accuracy, i.e. the probability that an ORS value matches the corresponding SRS value, and as reliability, i.e. the probability that a given SRS value is matched by the corresponding ORS value. Confidence intervals were calculated based on binomial and multinomial tests of accuracy.

A mixed model analysis of variance (ANOVA) was conducted to investigate if the ORS ratings are biased due to lighting conditions, i.e. if the raters expect the drivers to be sleepier during night-time and consequently provide higher ORS ratings in dark condition (and lower ORS ratings in the bright condition). *Condition* (day/night) was included as a fixed factor, the *SRS rating* (SRS 0 – SRS 2) was included as a fixed confounding factor and *Participant* (1–40) was included as a random factor. Multiple comparison post-hoc tests using Tukey's honestly significant difference criterion were used to investigate potential differences between individual groups. The significance level was set to 0.05.

If raters report low confidence in their ratings we expect that the match rate between ORS and SRS will be low. Similarly, we expect that when raters are confident in their ratings the match rate will be high. The relation between confidence level and match rate was investigated using a least squares best fit linear regression, with the expectation that the "ideal" correlation coefficient would be close to one.

Inter-individual differences between the raters were investigated using a similar approach to that of Wierwille and Ellsworth [17]. The covariance matrix of all raters' estimations of the 54 video clips was computed. The average was compared between all raters using Pearson's r correlation (i.e. the average of all the values above the main diagonal in the covariance matrix). Fisher z-transformation was used to make the correlation coefficients additive. Inter-individual differences were also assessed using Krippendorff's alpha using a difference function adapted to ordinal data [33].

3.3 Results

The overall average percentage of video segments where ORS and SRS matched was 41.5 % (confidence interval: 39.4–43.6 %). The relative frequency of matching segments can be seen in Fig. 2. A confusion matrix is provided in Table 2 and per-class accuracies and reliabilities, along with their confidence intervals are shown in Table 3. The ORS values (mean 0.92) was significantly lower than the SRS values (mean 1.00), $F_{(1,4317)} = 9.87$, $p < 0.01$.

There was a significant difference in ORS ratings due to *participant* ($F_{(39,2116)} = 2.07$, $p < 0.001$), *condition* ($F_{(1,2116)} = 7.99$, $p <= 0.01$) and *SRS rating* ($F_{(2,2116)} = 58.22$, $p < 0.001$). The post-hoc test revealed that higher ORS ratings were provided during night-time compared to daytime, and that increasing ORS ratings were accompanied by increasing SRS ratings. This means that the raters are influenced by lighting conditions. The distribution of ORS ratings per *condition* were ORS 0 (day 419, night 312), ORS 1 (day 373, night 488), and ORS 2 (day 288, night 279). A greater number of daytime video clips were scored as alert (ORS 0) while a higher number of nighttime driving video clips were

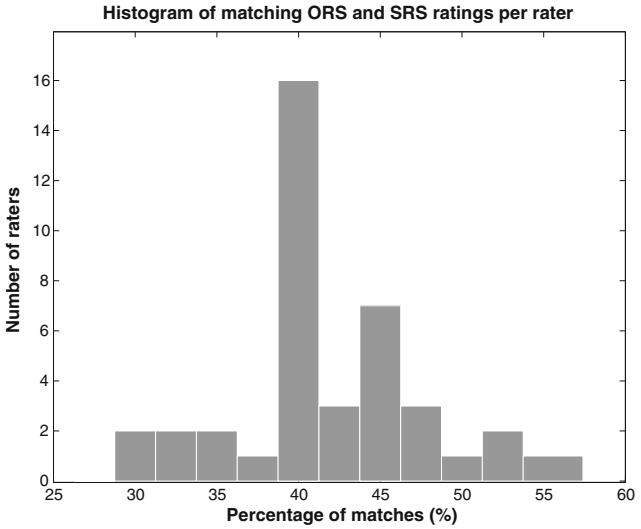

Fig. 2 Histogram of the raters' percentage of segments where ORS and SRS matched

scored as sleepy (ORS 1 or ORS 2). Of the three ORS levels, ORS 2 was the least scored and ORS 1 the most scored.

The confidence rating indicates how sure the raters were of that particular ORS estimation. In Fig. 3 the percentage of video segments where ORS and SRS matched are plotted against the confidence value. The slope of a linear regression line (least squares best fit) is 0.2 and the norm of residuals is 18.1, showing a very weak correspondence between matching rate and confidence. Boxplots of the confidence ratings for each of the 54 video clips are shown in Fig. 4. On the scale from 1 to 10, the median confidence rating is 6 for most clips, and only two clips are outside the range of the most common median value plus/minus 1 (i.e. confidence rating 5–7). This indicates that essentially no video clips are experienced as easier to rate.

Inter-individual differences between raters were determined as the average of the Pearson's r correlations between all raters, and was found to be 0.19. The correlation coefficients between raters varied a lot (standard deviation = 0.15). Krippendorff's alpha showed a slight agreement with $\alpha = 0.15$.

Table 2 Confusion matrix of ORS versus SRS

		ORS			
		0	1	2	Σ
SRS	0	339 (15.7 %)	262 (10.1 %)	118 (8.1 %)	719 (33.3 %)
	1	218 (12.1 %)	305 (14.1 %)	197 (13.6 %)	720 (33.3 %)
	2	174 (5.5 %)	294 (9.1 %)	252 (11.7 %)	720 (33.3 %)
	Σ	731 (33.8 %)	861 (39.9 %)	567 (26.3 %)	896 (41.5 %)

Table 3 Naïve per-class statistics for the matching of SRS and ORS

Level	Percentage match	95 % C.I.
	Accuracy	
Alert	46.4 %	42.7 % – 50.0 %
First signs of sleepiness	35.4 %	32.2 % – 38.7 %
Severe sleepiness	44.4 %	40.3 % – 48.6 %
	Reliability	
Alert	47.1 %	43.4 % – 50.9 %
First signs of sleepiness	42.4 %	38.7 % – 46.0 %
Severe sleepiness	35.0 %	31.4 % – 38.6 %

3.4 Discussion

ORS and SRS matched in 41 % of the 1-min video clips of drivers at various stages of sleepiness. This is only marginally better than chance (a matching percentage of 33.3 % is equivalent to someone picking ORS category at random for each video segment). This result is in agreement with the results from real-time ORS ratings, where 28 % of the changes in SRS were predicted by the observer [15].

Comparing the distribution of ORS and SRS demonstrates that for SRS 0 (alert), the raters' estimations matched SRS in 47 % of the video clips, and were completely unmatched (i.e. estimated ORS 2 for SRS 0) in only 17 % of the cases. The situation was worse for ORS 1 and ORS 2. One explanation is that it might be easier to identify an alert driver than a sleepy driver. Another explanation is that the raters may assume that the drivers are alert most of the time (an assumption which is certainly correct in a naturalistic dataset collected during normal driving). Such an assumption would systematically

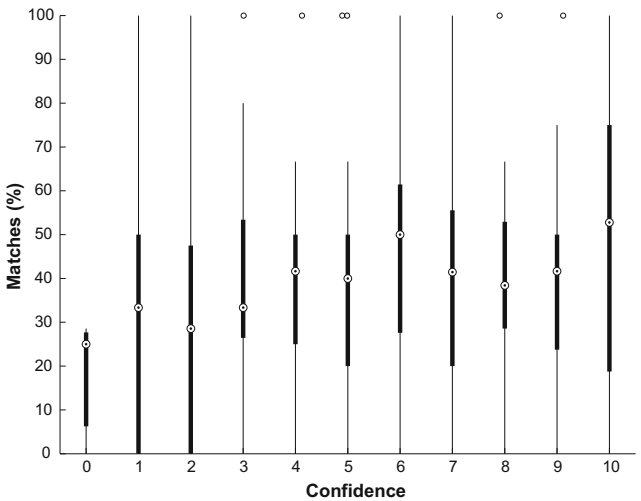

Fig. 3 Boxplot of the percentage of matching ORS and SRS ratings per confidence level. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, and the whiskers extend to $q_3 + 1.5(q_3 - q_1)$ and $q_1 - 1.5(q_3 - q_1)$, where q_1 and q_3 are the 25th and 75th percentiles, respectively. Outliers plotted individually

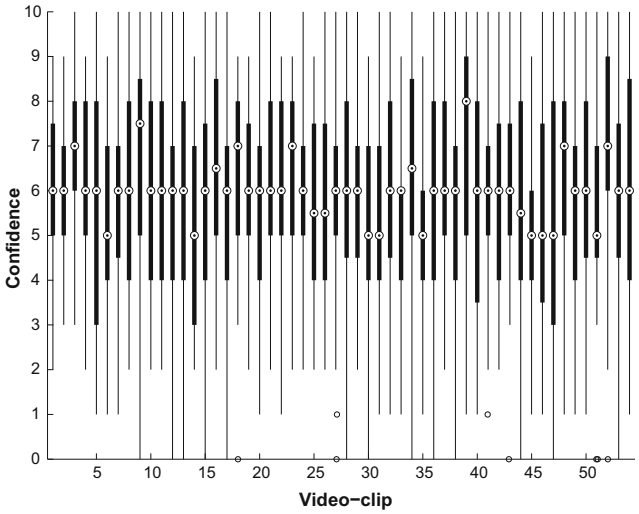

Fig. 4 Boxplots of the confidence in each rating grouped by video clip. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, and the whiskers extend to $q_3 + 1.5(q_3 - q_1)$ and $q_1 - 1.5(q_3 - q_1)$, where q_1 and q_3 are the 25th and 75th percentiles, respectively. Outliers plotted individually

underestimate sleepiness in our data set with an even distribution of high/medium/low SRS drives, as is also indicated by our results (the mean ORS rating is 0.92 whereas the mean SRS rating is 1.0).

The inter-rater agreement was very low in Test 1 with an average Pearson's r correlation of 0.19 and a Krippendorff's alpha of 0.15. This was unexpected given previous reports on high consistency and reliability both within and between raters [16, 17]. It is not clear whether this was due to inconsistent following of the ORS protocol, the ORS protocol in itself, inadequate video quality, or simply that it was difficult to rate sleepiness in a consistent manner.

Given the low match rate between ORS and SRS, it is interesting to see whether the raters are aware of their shortcomings. If so, this would be indicated in their confidence ratings. The expected matching rate, given perfect confidence estimates, would be a linear function increasing with the confidence level. However, Fig. 3 shows that this is not the case. Not only is it difficult to rate the sleepiness level of drivers based on video clips, it is also difficult for the raters to estimate the accuracy in their ratings.

4 Test 2

Test 1 raised a number of concerns about using ORS as a method to estimate driver sleepiness based on video recordings. A prerequisite for successful sleepiness scoring based on videos is whether it is at all possible to recognise a sleepy driver. To investigate this, a test was set up where the rating task was designed to be as easy as possible. In contrast to Test 1, Test 2 made use of colour-videos of higher quality, used a different camera position filming the driver from below instead of from the above, employed human factors experts as raters, and simplified the rating task to simply identify the video showing the sleepy condition in paired video clips showing the same driver in an alert and a sleepy condition.

4.1 Methods

Also Test 2 uses video data from an earlier experiment [34]. The videos were recorded from 18 bus drivers participating in a field experiment. The aim of the study was to investigate how a split shift with an early morning start influenced bus drivers' sleepiness level and driving performance during the late afternoon. Speed, GPS position and video films of the driver and of the forward view were continuously recorded throughout the driving session with a Video VBOX Pro (Racelogic Ltd., Buckingham, Great Britain). The video had a resolution of 720 × 576 pixels, where the face was an inset covering approximately one fourth of the picture. The driving experiment was carried out on a 23 km long route on public roads, which was driven three times. PVT was administered immediately before and after the driving session and between the second and third lap. In addition, physiology (EEG and EOG) and KSS (in 5 min intervals) was acquired. The results from the EEG measures are not used in this study.

The videos were selected based on KSS ratings, where the alert condition corresponded to KSS ≤ 4 and the sleepy condition corresponded to KSS ≥ 8. Unfortunately, only four of the 18 participants in the bus driver study matched this selection criteria. Thus, four matched pairs, comprising a set of eight 30-s video clips, were selected for this test. The pairs were matched in terms of driver and same type of road. The alert condition had an average (mean ± sd) blink duration of 113 ± 12 ms, a PVT reaction time of 348 ± 13 ms and a percentage of PVT lapses of 1.7 ± 1.1 %. In the sleepy condition, the average blink duration was 124 ± 15 ms, the PVT reaction time was 383 ± 42 ms and the percentage of PVT lapses was 8.0 ± 9.2 %. All videos were recorded in daytime during daylight conditions, removing the factor time of the day.

Ten participants (aged 38 ± 12 years, five women) were recruited to take part in the study as raters. All raters are part of the human factors group at the institute (VTI) and can be considered experts in the field of human behaviour assessment. After a short introduction the participants were shown video A, followed by video B, followed by video A and B next to each other. The participants were then asked to judge which of the two videos that showed the sleepy condition and how certain they were in their decision. The level of confidence was given in steps of 10 % on a scale from 0 %, meaning "no idea", to 100 %, meaning "absolutely certain". This procedure was repeated for the four video pairs.

4.2 Statistical analysis

The analyses involve descriptive statistics such as percentage correct classifications with corresponding mean confidence ratings. Inter-individual differences between the raters were investigated as in test 1 using Pearson's r correlation and Krippendorff's alpha.

4.3 Results

The overall average percentage of correctly assessed video segments was 35 %. The result from each video pair is presented in Table 4. Note that for two of the four video pairs, all raters identified the wrong video. Also provided in Table 4 are the average confidence ratings.

Inter-individual differences between raters were determined as the average of the Pearson's r correlations between all raters, and was found to be 0.35. Krippendorff's alpha showed a moderate agreement with $\alpha = 0.62$.

Table 4 Number of matches between ORS and SRS, including the raters' confidence in their ratings

Video pair	Match	Confidence (%)
1	9 of 10 (90 %)	51 ± 26
2	0 of 10 (0 %)	40 ± 24
3	5 of 10 (50 %)	42 ± 15
4	0 of 10 (0 %)	33 ± 20

4.4 Discussion

The inter-rater agreement was better in Test 2 than in Test 1 with an average Pearson's r correlation of 0.35 and a Krippendorff's alpha of 0.62. The difference in results between Test 1 and 2 may be due to a number of reasons. In Test 2, the video quality was better, the raters were more experienced, there was no confounding with lighting conditions, and the rating task was easier. However, the agreement was still very low.

The drivers were very sleepy in the sleepy condition. One driver experienced a short lapse of micro sleep and nearly hit the road barrier just a few minutes after the selected video clip (video pair 2). Another driver admitted that she provided KSS ratings of eight instead of nine since she was afraid that we would abort the trial in advance if we knew how sleepy she was (video pair 1). This is obviously just anecdotal evidence, but it is a clear indication that the pairs of videos contained footage of very varying levels of sleepiness. Despite our attempts to make an easy classification task, the raters disagreed on one of the video pairs and agreed on the "wrong" video in two of the four pairs. After discussions with the raters, the underlying reason for this result appears to be a difficulty of distinguishing boredom from sleepiness. The amount of data is very limited in Test 2, but the results consistently raise concerns about the validity of ORS.

5 General discussion

The notion that it is possible to assess the emotional state (e.g. sadness) or physiological state (e.g. sleepiness) of a fellow human with a single glance is widely held, but, the results of the current study shows how extremely difficult it can be to rate the sleepiness level of a driver using ORS.

Wierwille and Ellsworth [17] argue that it is impossible to know the extent to which the raters are rating the "true sleepiness level" and therefore chose not to do such comparisons. Instead, they chose to use consensus amongst the observers as a reference measure. Given the poor correspondence between SRS and ORS found in our test and the difficulty of distinguishing boredom from sleepiness, using consensus amongst observers might be a dangerous way of determining the "truth". Wiegand et al. [16] instead chose to use experienced raters as a reference measure of sleepiness and found agreement with non-expert ORS to be satisfactory. Again, given the poor agreement between ORS and SRS in our tests, also amongst the expert raters, expert judgements may be misleading when assessing sleepiness.

There may be individual differences in subjectively experienced levels of sleepiness and in behavioural signs of sleepiness, meaning that SRS and ORS may represent different constructs of sleepiness. A further indication that this may be the case is that the percentage of matching ORS/SRS values did not increase as a function of confidence in the ratings. As the list of agreeing measures of sleepiness grows longer (subjective ratings, physiology, performance) while post-hoc ORS stands alone on the disagreeing side, it may be worth considering that ORS is measuring a different construct of sleepiness. Perhaps ORS is rather an indicator of task related fatigue, boredom and sleepiness in a broader sense. There are too many limitations in the current data material to draw such conclusions, but further studies should be conducted to confirm or reject this suggestion.

There are some controversies when it comes to the influence of video clip duration and image quality. For starters, behavioural signs of sleepiness such as yawning and nodding are rather infrequent events that may not be captured in a short recording. Ratings of sleepiness based on such events will obviously become more reliable with longer video clips. On the other hand, fatigue has been successfully rated in photographs based on hanging eyelids, red eyes, swollen eyes, darker circles under the eyes, paler skin, more wrinkles/fine lines, droopy corners of the mouth and glazed eyes [24], indicating that the image quality is perhaps even more important than the duration of the clip. However, if video quality was the

only issue, we would have expected better results for real-time ORS [15]. In real-time ORS the driver is observed directly, thus providing the ultimate "image" quality, and also indirectly taking driving performance into account. Also, in contrast to the Sundelin [24] study, our sleepy drivers actively fought to remain awake, which may mask some of the observable signs of sleepiness. Higher video quality may solve the issue with poor inter-rater agreement. As the visual cues becomes more consistent and easier to extract, the ratings are also most likely to become more consistent. Whether improved visual cues are directly related to sleepiness and SRS is however unclear. There will be an impact on the match rate between ORS and SRS, but it is not clear if the match rate will increase or decrease. As already indicated, some of these cues may rather be signs of boredom or fatigue, and this may be misleading when it comes to sleepiness assessment.

A limitation with this research is that the videos and the data used in both Test 1 and Test 2 comes from earlier experiments that had little to do with the main questions presented here. An implication is, for example, that very few video pairs could be used in Test 2 due to the selection criterion that the same driver should have experienced both high and low KSS values. Another limitation is the level of training given to the untrained raters in Test1. Even though the amount of training is similar to what we would give to an actual "scientific" ORS rater at our institute, a "scientific" rater would still be better equipped due to years of experience gained by riding along in the vehicle as a test leader in driver sleepiness experiments. Since such a "scientific" rater would have a qualitatively different background, this could be a serious shortcoming in Test 1. However, the "scientific" raters performed poorly in Test 2, so we are not sure that this is such a big limitation in practice. Additional studies, replicating and verifying the results from this paper, should be carried out.

Much research is currently devoted to developing automatic image processing systems capable of determining the level of sleepiness based on characteristics such as facial tone, slow eyelid closure, rubbing, yawning and nodding [35–39]. According to Vural et al. [40], the ten facial actions that are most predictive of sleepiness are increased blink/eye closure, elevated outer brow raise, increased frown, chin raise, more nose wrinkle, less smiling, tightened eye lid, less compressed nostrils, less lowering of eye brows and less jaw drop. It would be very interesting to compare the outcome of these automatic facial emotion recognition systems to self-rated and observer rated sleepiness, and also to investigate which features that are used by the automatic system compared to the human observer.

In conclusion, the results from this study indicate that video-based ORS ratings of driver sleepiness correspond poorly with self-reported sleepiness. In this study, the raters were even unable to pick out a sleepy driver in a paired comparison where they know that one of the videos showed a sleepy driver. Further investigations should be carried out to evaluate the reliability of ORS and to inform on future use of this technique.

Acknowledgments This study was funded by the competence centre *Virtual Prototyping and Assessment by Simulation* (ViP) and by the *Vehicle and Traffic Safety Centre* (SAFER) at Chalmers University of Technology. The authors are very thankful to Ashleigh Filtness, Queensland University of Technology, for fruitful discussions during the writing of this paper.

References

1. Hanowski RJ, Wierwille WW, Dingus TA (2003) An on-road study to investigate fatigue in local/short haul trucking. Accid Anal Prev 35(2):153–160. doi:10.1016/S0001-4575(01)00098-7
2. Dingus TA, Klauer SG, Neale VL, Petersen A, Lee SE, Sudweeks J, Perez MA, Hankey J, Ramsey D, Gupta S, Bucher C, Doerzaph ZR, Jermeland J, Knipling RR (2006) The 100-car naturalistic driving study, phase II, results of the 100-car field experiment. NHTSA, Washington, DC
3. Hanowski RJ, Blanco M, Nakata A, Hickman JS, Schaudt WA, Fumero MC, Olson RL, Jermeland J, Greening M, Holbrook GT, Knipling RR, Madison P (2008) The drowsy driver warning system field operational test, data collection final report. NHTSA
4. Klauer SG, Dingus TA, Neale VL, Sudweeks J, Ramsey D (2006) The impact of driver inattention on near-crash/crash risk: an analysis using the 100-car naturalistic driving study data. NHTSA, Washington DC
5. Putilov AA, Donskaya OG, Verevkin EG (2012) Quantification of sleepiness through principal component analysis of the electroencephalographic spectrum. Chronobiol Int 29(4):509–522. doi:10.3109/07420528.2012.667029
6. Cluydts R, DE Valck E, Verstraeten E, Theys P (2002) Daytime sleepiness and its evaluation. Sleep Med Rev 6(2):83–96. doi:10.1053/smrv.2002.0191
7. Horne JA, Baulk SD (2004) Awareness of sleepiness when driving. Psychophysiology 41(1):161–165. doi:10.1046/j.1469-8986.2003.00130.x
8. Lal SKL, Craig A (2002) Driver fatigue: electroencephalography and psychological assessment. Psychophysiology 39(3):313–321
9. Campagne A, Pebayle T, Muzet A (2005) Oculomotor changes due to road events during prolonged monotonous simulated driving. Biol Psychol 68(3):353–368. doi:10.1016/j.biophyscho.2004.05.003
10. Anund A, Kecklund G, Kircher A, Tapani A, Åkerstedt T (2009) The effects of driving situation on sleepiness indicators after sleep loss: a driving simulator study. Ind Health 47(4):393–401
11. Ingre M, Åkerstedt T, Peters B, Anund A, Kecklund G (2006) Subjective sleepiness, simulated driving performance and blink duration: examining individual differences. J Sleep Res 15(1):47–53. doi:10.1111/j.1365-2869.2006.00504.x
12. Philip P, Sagaspe P, Moore N, Taillard J, Charles A, Guilleminault C, Bioulac B (2005) Fatigue, sleep restriction and driving

performance. Accid Anal Prev 37(3):473–478. doi:10.1016/j.aap.2004.07.007
13. Åkerstedt T, Ingre M, Kecklund G, Anund A, Sandberg D, Wahde M, Philip P, Kronberg P (2010) Reaction of sleepiness indicators to partial sleep deprivation, time of day and time on task in a driving simulator–the DROWSI project. J Sleep Res 19(2):298–309. doi:10.1111/j.1365-2869.2009.00796.x
14. Maldonado CC, Bentley AJ, Mitchell D (2004) A pictorial sleepiness scale based on cartoon faces. Sleep 27(3):541–548
15. Anund A, Fors C, Hallvig D, Akerstedt T, Kecklund G (2013) Observer rated sleepiness and real road driving: an explorative study. PLoS One 8(5):e64782
16. Wiegand DM, McClafferty J, McDonald SE, Hanowski RJ (2009) Development and evaluation of a naturalistic observer rating of drowsiness protocol. The National Surface Transportation Safety Centre for Excellence
17. Wierwille WW, Ellsworth LA (1994) Evaluation of driver drowsiness by trained raters. Accid Anal Prev 26(5):571–581
18. Åkerstedt T, Anund A, Axelsson J, Kecklund G (2014) Subjective sleepiness is a sensitive indicator of insufficient sleep and impaired waking function. J Sleep Res 23(3):240–252. doi:10.1111/jsr.12158
19. Rogé J, Pebayle T, Muzet A (2001) Variations of the level of vigilance and of behavioural activities during simulated automobile driving. Accid Anal Prev 33(2):181–186. doi:10.1016/S0001-4575(00)00029-4
20. Barr LC, Yang CYD, Hanowski RJ, Olson R (2011) An assessment of driver drowsiness, distraction, and performance in a naturalistic setting. Federal Motor Carrier Safety Administration
21. Klauer SG, Guo F, Sudweeks J, Dingus TA (2010) An analysis of driver inattention using a case-crossover approach on 100-car data: final report. NHTSA, Washington DC
22. Blanco M, Bocanegra JL, Morgan JF, Fitch GM, Medina A, Olson R, Hanowski RJ, Daily B, Zimmermann RP, Howarth HD, DI Domenico TE, Barr LC, Popkin SM, Green K (2009) Assessment of a drowsy driver warning system for heavy-vehicle drivers: final report. NHTSA
23. Axelsson J, Sundelin T, Ingre M, VAN Someren EJW, Olsson A, Lekander M (2010) Beauty sleep: experimental study on the perceived health and attractiveness of sleep deprived people. BMJ 341. doi:10.1136/bmj.c6614
24. Sundelin T, Lekander M, Kecklund G, VAN Someren EJ, Olsson A, Axelsson J (2013) Cues of fatigue: effects of sleep deprivation on facial appearance. Sleep 36(9):1355–1360. doi:10.5665/sleep.2964
25. Drummond SPA, Bischoff-Grethe A, Dinges DF, Ayalon L, Mednick SC, Meloy MJ (2005) The neural basis of the psychomotor vigilance task. Sleep 28(9):1059–1068
26. McKinley RA, McIntire LK, Schmidt R, Repperger DW, Caldwell JA (2011) Evaluation of eye metrics as a detector of fatigue. Hum Factors J Human Factors Ergon Soc 53(4):403–414. doi:10.1177/0018720811411297
27. Wright KP, Hull JT, Czeisler CA (2002) Relationship between alertness, performance, and body temperature in humans. Am J Physiol Regul Integr Comp Physiol 283:R1370–R1377
28. Jammes B, Sharabty H, Esteve D (2008) Automatic EOG analysis: a first step toward automatic drowsiness scoring during wake-sleep transitions. Somnologie - Schlafforschung und Schlafmedizin 12(3):227–232. doi:10.1007/s11818-008-0351-y
29. Loh S, Lamond N, Dorrian J, Roach G, Dawson D (2004) The validity of psychomotor vigilance tasks of less than 10-minute duration. Behav Res Methods Instrum Comput 36(2):339–346
30. Åkerstedt T, Gillberg M (1990) Subjective and objective sleepiness in the active individual. Int J Neurosci 52(1–2):29–37
31. Ahlstrom C, Nystrom M, Holmqvist K, Fors C, Sandberg D, Anund A, Kecklund G, Akerstedt T (2013) Fit-for-duty test for estimation of drivers' sleepiness level: eye movements improve the sleep/wake predictor. Transp Res C Emerg Technol 26:20–32. doi:10.1016/j.trc.2012.07.008
32. Fors C, Ahlström C, Sörner P, Kovaceva J, Hasselberg E, Krantz M, Grönvall J-F, Kircher K, Anund A (2011) Camera-based sleepiness detection: final report of the project SleepEYE. VTI, Linköping
33. Krippendorff K (2013) Content analysis: an introduction to its methodology. Sage, Thousand Oaks
34. Anund A, Kecklund G, Fors C, Ihlström J, Söderström B (2014) Bussförares arbetstider kopplat till trötthet. VTI report R830. VTI, Linköping, Sweden
35. Bergasa LM, Nuevo J, Sotelo MA, Barea R, Lopez ME (2006) Real-time system for monitoring driver vigilance. IEEE Trans Intell Transp Syst 7(1):63–77. doi:10.1109/TITS.2006.869598
36. Boyraz P, Acar M, Kerr D (2008) Multi-sensor driver drowsiness monitoring. Proc IME D J Automob Eng 222(D11):2041–2062. doi:10.1243/09544070JAUTO513
37. Flores MJ, Armingol JM, DE LA Escalera A (2010) Driver drowsiness warning system using visual information for both diurnal and nocturnal illumination conditions. EURASIP J Adv Signal Process. doi:10.1155/2010/438205
38. Flores MJ, Armingol JM, DE LA Escalera A (2010) Real-time warning system for driver drowsiness detection using visual information. J Intell Robot Syst 59(2):103–125. doi:10.1007/s10846-009-9391-1
39. Jimenez-Pinto J, Torres-Torriti M (2012) Face salient points and eyes tracking for robust drowsiness detection. Robotica 30(5):1–11. doi:10.1017/S0263574711000749
40. Vural E, Cetin M, Ercil A, Littlewort G, Bartlett M, Movellan J (2007) Drowsy driver detection through facial movement analysis. In: Lew M, Sebe N, Huang T, Bakker E (eds) Human–computer interaction, vol 4796. Lecture Notes in Computer Science. Springer, Berlin Heidelberg, pp. 6–18. doi:10.1007/978-3-540-75773-3_2