Detecting Short Periods of Elevated Workload: A Comparison of Nine Workload Assessment Techniques

Willem B. Verwey and Hans A. Veltman
TNO Human Factors Research Institute

The present experiment tested the merits of 9 common workload assessment techniques with relatively short periods of workload in a car-driving task. Twelve participants drove an instrumented car and performed a visually loading task and a mentally loading task for 10, 30, and 60 s. The results show that 10-s periods of visual and mental workload can be measured successfully with subjective ratings and secondary task performance. With respect to longer loading periods (30 and 60 s), steering frequency was found to be sensitive to visual workload, and skin conductance response (SCR) was sensitive to mental workload. The results lead to preliminary guidelines that will help applied researchers to determine which techniques are best suited for assessing visual and mental workload.

In recent years, major research programs in Europe, the United States, and Japan have boosted the development of information systems in road traffic. In Europe, these systems are referred to as transport telematics, but in the United States the term intelligent transport systems (ITS) is more common. It is generally expected that transport telematic applications will change the nature of car driving considerably in the near future. A vast array of telematics systems is currently being considered. Some well-known examples are route guidance systems, anticollision systems, systems for monitoring driver status, and parking advice systems. However, installing even a subset of these telematic systems, many of which are aimed at supporting the driver and increasing traffic safety, may lead to distraction and overload of the driver and will result in increased accident rates. Therefore, care should be taken to design telematics systems according to a driver's needs and capabilities (e.g., Michon, 1993). This implies that the human factors specialist should have detailed knowledge of the advantages and disadvantages of the existing techniques used for assessing driver workload.

In a review of workload assessment methodologies, O’Donnell and Eggemeier (1986) proposed five criteria for selecting workload assessment techniques. Two major criteria were sensitivity (whether the technique discriminates between levels of workload) and diagnosticity (whether the technique distinguishes between types of workload). The purpose of the present article is to compare a series of nine common workload assessment techniques to relatively short periods of elevated workload in driving. Special emphasis is given to the sensitivity and diagnosticity of these techniques.

Several studies have compared workload assessment techniques. These studies have yielded different results with respect to sensitivity of these techniques. For example, Casali and Wierwille (1984) compared workload assessment techniques in a flight simulator study. Hicks and Wierwille (1979) did the same in a driving simulator study. In both studies, subjective rating scales and primary task performance were able to quantify workload of the primary task. However, secondary task performance was a poor workload indicator in the
second study. Neither study showed an effect of workload on heart rate (HR) and HR variability, but still other studies did show that HR measures were sensitive indicators for workload caused by traffic jams and bad weather (Vivoli, Bergomi, Rovesti, Carrozzi, & Vezzosi, 1993) and workload caused by telephoning while driving (Brookhuis, de Vries, & de Waard, 1991). The outcomes of these studies illustrate that there are marked sensitivity differences among techniques in different tasks and in different driving environments.

A cause for the differences between studies may be that the techniques differ with respect to their sensitivity to rapid fluctuations in workload level. For example, if workload is assessed in a task involving short workload peaks and a technique is used that is highly sensitive to peaks, the study will show a generally high level of workload. When a technique is used that is not sensitive to peaks, the conclusion will be that workload is low. However, in tasks involving only gradual changes in workload, both techniques may indicate similar workload levels.

The sensitivity of workload assessment techniques to rapid fluctuations in workload has received some attention (e.g., Knowles, 1963; Wierwille, 1981), but empirical studies are scarce. Antin and Wierwille (1984) compared the sensitivity of six workload assessment techniques with respect to the workload induced by two laboratory tasks. Level of workload changed rapidly in these tasks. Primary task performance and subjective estimates yielded the best results, but the duration of workload elevations was not systematically manipulated. It remains unclear how long workload increments must last for them to be detected by the various techniques. Moreover, it is unclear to what extent these results can be extrapolated to the driving environment, which is the task environment of primary interest in this article.

In some models, there are various types of workload. For example, Wickens (1984b) distinguished workload on perceptual and response modalities, on cognitive coding mechanisms, and on processing stages. In line with this model, Verwey (1993) demonstrated that during driving visual and mental workload may diverge. Visual workload was relatively high in situations involving changes in course, such as at turns and on curves. Mental workload was affected much less in these situations. This shows that a workload assessment technique cannot be selected without taking the question at hand into account. For example, when workload is used to indicate error probability, another technique is required than when exploring the possibilities to add a task to a task environment. In the latter case, a technique is required to allow the researcher to diagnose types of elevated workload. Estimating error probability requires a technique sensitive to all types of workload. However, little is known about the diagnosticity of the various workload assessment techniques, that is, to what extent the various techniques are sensitive to one or another type of workload.

The aim of the present study was to compare the sensitivity and diagnosticity of a series of workload assessment techniques for short periods of elevated visual and mental workload. Experienced drivers operated an instrumented car on a straight four-lane freeway. This task involved the basic elements of driving, that is, lane keeping and controlling speed. To increase workload in a controlled manner, we added a visually loading task and a mentally loading task to the driving task. These are called loading tasks. Both loading tasks lasted for 10, 30, or 60 s. Even though peaks shorter than 10 s can occur in driving, workload elevations of 10 s are still considered peaks. Periods of 30 and 60 s of elevated workload were introduced to obtain an indication of the sensitivity and diagnosticity of the workload assessment techniques for longer periods of high workload.

Workload was assessed with nine different techniques: two primary task measures, one secondary task measure, four physiological measures, and two subjective measures. Primary task measures were driving speed and the intervals between successive steering movements. Speed has been found to decrease with increasing workload (Harms, 1986; Verwey & Janssen, 1989). Steering frequency increased under such conditions (Antin, Dingus, Hulse, & Wierwille, 1990; MacDonald & Hoffman, 1980; Verwey, 1991).

A fairly common technique for measuring workload is secondary task performance. The rationale is that performance on a low-priority task reflects the workload induced by a concurrent high-priority task (e.g., Noy, 1987). To ensure that participants acknowledge this priority and continue doing so, the researcher should repeatedly tell them to perform the secondary task only if the primary task performance does not deteriorate.
(Wickens, 1984a). The diagnosticity of the secondary task technique depends on the characteristics of the secondary task being used (O'Donnell & Eggemeier, 1986; Wickens, 1984b). A secondary task generating substantial visual workload is diagnostic for the visual workload of the primary task. A mentally loading secondary task is diagnostic for mental workload.

Various physiological measures for workload were used: two measures derived from the electrocardiogram, including interbeat interval (IBI) and HR variability, the interval between successive eyeblinks, and the interval between skin conductance responses (SCRs). An increase in workload would result in a decrease of IBI and HR variability because operators increase their level of arousal or mental effort in order to accomplish their goals (Gaillard & Wientjes, 1994; Hockey, 1986).

Eyeblink intervals were measured because people are known to suppress the eyeblink reflex when they have to process visual information (Stern, Boyer, & Schroeder, 1994; Veltman & Gaillard, 1996; Wilson & Fisher, 1990). The extent to which mental workload affects eyeblinks is not entirely clear. There are indications that eyeblinks do not respond to mental task demands (Casali & Wierwille, 1984; Wierwille & Eggemeier, 1993). This suggests that eyeblinks are diagnostic with respect to visual workload.

A measure that is traditionally associated with workload, and especially with arousal states accompanied by effort and emotions, is the spontaneous electrodermal activity or SCR (e.g., Edelberg, 1972). In a classic study, Taylor (1964) found a relationship between the number of SCRs and accident rate, number of turns per distance unit, and driving speed. Zeier (1979) observed that driving in heavy city traffic with a manual transmission yielded more SCRs than driving with an automatic transmission, which would suggest that SCR indexes activation of the sympathetic nervous system.

Because physiological measures of workload are often associated with arousal and emotions, and consequently effort, one might expect that SCR, IBI, HR variability, and eyeblinks are closely related and have limited correlations with performance-based measures of workload (Casali & Wierwille, 1984; Hart & Wickens, 1990). The reason is that effort and task performance need not coincide. It is important to find whether these physiological measures will be sensitive to short workload periods in driving, as many previous studies showed either limited or no sensitivity of physiological measures (Wierwille, 1979).

Finally, workload was assessed by two subjective rating scales: the Subjective Workload Assessment Technique (SWAT) and the Rating Scale Mental Effort (RSME). The SWAT (Reid & Nygren, 1988) is based on the premise that workload is a multidimensional concept that can be assessed by asking participants for their workload in terms of time pressure, mental effort, and psychological stress. Assessing workload on these three dimensions would increase the sensitivity and the diagnosticity of this technique. In addition, the SWAT takes individual differences into account. Participants are requested to perform a card-sorting procedure that identifies relative sensitivity among these dimensions. The RSME (Zijlstra, 1993) is a univariate rating scale. This scale ranges from 0 to 150 and has nine descriptive indicators along its axis (e.g., not effortful and awfully effortful). Validation of this technique ensures that the meaning of the verbal labels are the same for different people.

A multidimensional and a univariate scale were included because there has been a dispute about the usefulness for rating scales to have more than one dimension. Comparative evaluations of subjective scales conducted in a wide range of tasks showed little advantage for the multidimensional scales (see Wierwille & Eggemeier, 1993, for a review). In addition, there is mounting evidence that taking individual differences into account by way of the sorting task adds little or no sensitivity. However, because this has been found with rather homogeneous participant groups such as pilots, the question of whether individual differences are important in more heterogenous groups, such as car drivers remains.

In summary, the present study compared nine workload assessment techniques in the driving environment with respect to their sensitivity and their diagnosticity to 10-, 30-, and 60-s periods of elevated workload. These techniques encompassed two driving parameters (speed and steering intervals), secondary task performance (visual detection), two rating scales (SWAT and RSME), and four physiological measures (IBI, HR variability, eyeblink intervals, and SCR). Detailed insight into the merits of these workload assessment techniques for the driving task will aid in the
design of future studies of driver workload and be of theoretical interest for understanding workload.

Method

Driving and Loading Tasks

The experimental route consisted of 40-km, four-lane freeway connecting two Dutch cities, Amersfoort and Apeldoorn. This road is characterized by relatively stable traffic with little or no congestion. It took 25 to 30 min to drive the experimental route in a single direction. Participants were instructed to drive 80 to 90 kph in the right lane and to avoid overtaking.

To impose workload for predefined periods of time on top of the workload caused by lane and speed keeping, participants carried out loading tasks for 10, 30, or 60 s. Peaks below 10 s seemed too short to be registered by the present techniques. The 30- and 60-s periods were included to assess sensitivity and diagnosticity to longer periods of increased workload as well. One loading task used a set of four high luminance light-emitting diodes (LEDs) placed in a rectangular form. Only three of the LEDs were switched on at the time. The resulting shape changed each 1 to 2 s. There was a probability of .25 that the LEDs formed an L shape. Participants were to detect and count the number of times the L shape was shown. This task was called the L-counting task.

The second loading task was the continuous memory task (CMT). The CMT requires participants to count each of a number of target letters in a stream of spoken letters (Boer & Jorna, 1987). In the present study, the targets were A, B, and C. Targets had a higher presentation probability than nontargets; for each nontarget probability was .03; for each target this probability was .14. In effect, there was a probability of .42 that a target was presented. The interval between onset of succeeding messages varied randomly between 2 and 4 s. Presentation time was 750 ms. Participants were urged to give this task low priority relative to all other tasks, including the loading task. Care was taken to make sure that there was no synchronization between stimulus presentation in the secondary task and in the loading tasks. The secondary task continued irrespective of whether a loading task was carried out.

A third group of workload assessment techniques involved the SWAT and RSME rating scales. The SWAT is based on participant ratings from 1 to 3 on time pressure, effort, and stress. Individual differences are taken into account by integrating the results of a card-sorting procedure that identifies relative weight of each of these dimensions. The RSME is a one-dimensional scale with ratings between 0 and 150. The scale has nine descriptive indicators along its axis (e.g., 2 corresponds to not effortful, 58 to rather effortful, and 113 to awfully effortful). It is designed so as to minimize individual differences.

Finally, there were four different physiological measures for workload: IBI, HR variability, intervals between successive eyeblinks, and intervals between successive SCRs. IBI and HR variability

Workload Assessment Techniques

Nine techniques were used to assess the workload of the driver. Two involved primary task measures, that is, indicators of driving performance that are known to be affected by driver workload. The first primary task measure was driving speed. Second, a measure of steering frequency called steering reversal rate (SRR) was assessed by analyzing the average time between successive wheel movements. A movement was defined as a change from a negative (clockwise movement) to a positive (counterclockwise movement) rotational velocity, given that the positive rotational velocity exceeded 1° per second.

The secondary task technique for assessing workload (not to be confused with the loading tasks) was a visual detection task. It required participants to say the Dutch equivalent of "yes" (i.e., "ja") upon detecting a target stimulus (two random punctuation marks) on a dashboard-mounted display. To prevent peripheral detection of stimuli, we presented a neutral stimulus (called "gg") between the target stimuli. The intervals between onset of succeeding messages varied randomly between 2 and 4 s. Presentation time was 750 ms. Participants were urged to give this task low priority relative to all other tasks, including the loading task. Care was taken to make sure that there was no synchronization between stimulus presentation in the secondary task and in the loading tasks. The secondary task continued irrespective of whether a loading task was carried out.
were derived from the electrocardiogram (ECG) and eyblinks from the electrooculogram (EOG).

**Apparatus and Data Collection**

The experiment was carried out in an instrumented Volvo 240 station wagon with dual controls. This car contained an IBM 486 personal computer and various auxiliary apparatus for measuring driving behavior and generating stimuli. In this study, speed, steering wheel position, and stimuli presented were registered at a sample rate of 10 Hz.

The readings for ECG, EOG, and SCR were digitally recorded at 100 Hz with Codas software (DATAQ Instruments, Akron, OH) by means of silver/silver chloride (Ag/AgCl) electrodes. We used electrodes placed on drivers' chests for the ECGs. The HR variability coefficient was computed by dividing the standard deviation of the IBIs by the average IBI (van Dellen, Aasman, Mulder, & Mulder, 1983). The EOG was measured with electrodes above and below the right eye and recorded with an AC-coupled amplifier ($\tau = 3$ s). The derivative of the EOG signal was used for the detection of eyblinks. A blink was defined as a valley followed by a peak in the differentiated signal within less than 0.2 s and the amplitude of the peaks and valleys had to be at least 50% of the standard deviation of the entire signal. The SCR was measured with a Wheatstone bridge and amplified by an AC-coupled amplifier ($\tau = 10$ s). The electrodes were attached to the palm of the participant's left hand about 4 cm apart. Fluctuations in SCR were detected by Codas software and checked visually.

The CMT stimuli were uttered by a digitized female voice and were clearly audible under all driving conditions. L-counting stimuli were presented by LEDs mounted in a box at the left side of the steering wheel. The LEDs were protected from glare by a cardboard cover. The visual angle between the normal fixation point on the road ahead and the LEDs was about 27.0° horizontally and 20.0° vertically. Again, a black cardboard shield was placed around this display. The height of the two digit stimuli on the screen was about 0.9° and the width was 0.7° per digit. The L-counting LEDs and the secondary task plasma display could not be seen at a single glance.

**Participants**

Twelve men participated in the present study. They were all between the ages of 23 and 50, had their licenses for more than 5 years, and had driven more than 10,000 km per year in the last 5 years.

**Procedure**

Participants were familiarized with the aim of the study, the instrumented car, the SWAT and RSME, the secondary task, and the two loading tasks. They were instructed to rate workload on the SWAT and RSME only with respect to the preceding period of loading task exposure. It was explicitly stated before and during the experiment that the secondary task (the visual detection task) was relatively unimportant, whereas the loading tasks (the L-counting task and the CMT) were extremely important and that target stimuli should not be missed. Furthermore, participants were repeatedly instructed to prevent the driving task from deteriorating when performing other tasks. Then, the SWAT card-sorting task was carried out, and electrodes for ECG, EOG, and SCR registration were attached.

First, physiological baseline conditions were registered for 15 min in the car while the participants were reading a newspaper. While driving toward the experimental route, the participants practiced the loading tasks and the secondary task for about 10 min. Order of the seven experimental conditions was counterbalanced across participants. For safety reasons, a licensed driver-training instructor who was able to take over with dual controls acted as experimenter.

Participants were warned by an auditory signal when L-counting or the CMT started. The 10-, 30-, and 60-s loading task blocks were presented in a random order. Response accuracy in the loading tasks was recorded at the end of each block by the experimenter after accuracy was verified by the technician. When a response was incorrect, the participant was informed immediately. To avoid
mishearing the vocal responses, participants wore a lightweight head-mounted microphone that was connected to a headphone used by the technician.

**Data Analysis**

Independent variables were loading task (no-loading task control, L-counting, and continuous memory task), duration (10, 30, and 60 s), and period (loading task vs. no-loading task). Figure 1 shows that each loading task block was surrounded by two 60-s rest periods. During these periods, participants merely drove the car. When appropriate, they continued with the secondary task. Physiological and driving measures were assessed in these 60-s blocks too, but, as subjective ratings were given by the participants at the start of these blocks, these measures were analyzed only for the last 30 s. The average workload assessed in the two rest periods surrounding each loading task block was used as local control for that block.

The SWAT results were analyzed by adding the ratings on the three scales and neglecting the results of the card-sorting procedure. Moreover, the relative weight of each dimension was derived from the card sorting and combined with the ratings at the three scales in three different ways. The weights of the dimensions were computed (a) for the entire group, (b) for groups of prototypical participants, and (c) for individual participants. The card-sorting results showed that the group solution (option a above) was allowed because the coefficient of concordance, as computed by the analysis package provided with the SWAT, amounted to .82. The recommended minimal coefficient of concordance for group solutions is .78. Separate analyses on the SWAT workload results confirmed this. For this reason, only the analysis that was based on the solution for the entire participant group are reported below. The RSME was analyzed by subjecting the raw ratings to an analysis of variance (ANOVA).

Analysis of the SCR signal was based on the weighted average interval between SCRs during the periods of elevated workload. The last SCR before and the first SCR after this period were also included in this analysis, as they marked intervals that partly overlapped with the loading task periods.

**Results**

**Performance on the Loading Tasks: L-Counting and CMT**

On average, 5.3% of the L-counting targets were missed. This percentage was higher when the secondary task was carried out, 8.4% versus 2.3%; $F(1, 11) = 13.2, p < .01$. The CMT performance was indexed by the sum of the absolute differences between the number of detected and the number of presented target letters. Across participants, the number of detected targets deviated 6.4% from the number of presented targets.

**General Workload Indicators**

Statistical significance of the effects of the loading tasks are indicated in Table 1 for each dependent variable. Comparisons are made between both loading task conditions and the control condition in the first $F/p$ column. The second $F/p$ column in Table 1 shows significance of the local differences by comparing the dependent variables obtained during actual execution of the loading tasks and those measured during the rest periods before and after them. The third $F/p$ column gives rough indications of the sensitivity of the dependent variables to the duration of elevated workload. These results indicate that steering frequency, secondary task performance, SWAT, RSME, and, to a lesser degree, HR variability, and SCR were sensitive to the workload induced by the loading tasks. In addition, levels of workload indicated by secondary task performance, SWAT, RSME, and steering frequency were affected by workload duration.

**Driving Performance**

Participants drove 84.6 kph on the average, but, as indicated in Table 1, speed was not affected by
Table 1
F Values and Levels of Statistical Significance With Respect to Global Loading Task Effects, Local Loading Task Effects, and Duration Effects for Each of Nine Dependent Variables

<table>
<thead>
<tr>
<th>Task</th>
<th>Control vs. L-counting/CMT</th>
<th>Control vs. L-counting/CMT x Period</th>
<th>Control vs. L-counting/CMT x Duration*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F(1, 11)</td>
<td>p</td>
<td>F(1, 11)</td>
</tr>
<tr>
<td>Speed</td>
<td>0.2</td>
<td>&gt;.10</td>
<td>1.2</td>
</tr>
<tr>
<td>Steering</td>
<td>18.5</td>
<td>&lt;.001</td>
<td>6.9</td>
</tr>
<tr>
<td>Secondary task</td>
<td>38.9</td>
<td>&lt;.001</td>
<td>33.1</td>
</tr>
<tr>
<td>SWAT</td>
<td>30.9</td>
<td>&lt;.001</td>
<td>28.3</td>
</tr>
<tr>
<td>RSME</td>
<td>43.2</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>Interbeat interval</td>
<td>0.2</td>
<td>&gt;.10</td>
<td>2.7</td>
</tr>
<tr>
<td>Heart rate variability</td>
<td>3.8</td>
<td>&lt;.10</td>
<td>1.7</td>
</tr>
<tr>
<td>Eyeblinks</td>
<td>0.6</td>
<td>&gt;.10</td>
<td>0.3</td>
</tr>
<tr>
<td>Skin conductance response</td>
<td>4.1</td>
<td>&lt;.10</td>
<td>8.6</td>
</tr>
</tbody>
</table>

Note. SWAT = Subjective Workload Assessment Technique; RSME = Rating Scale Mental Effort; CMT = continuous memory task.
*These planned comparisons involved 10 versus 60 s only. Control versus CMT: L-counting (the number of L-shaped light-emitting diodes) excluded because eyeblinks and vertical eye movements could not be distinguished.

The loading tasks. The secondary (workload assessment) task did not affect speed, F(1, 11) = 1.6, p > .20.

Steering intervals were generally longer during the loading task period than in the no-loading task period (see Figure 2), F(1, 11) = 51.4, p < .001. However, Figure 2 suggests that steering intervals decreased more with L-counting than with the control task and the CMT. This was confirmed by pairwise planned comparisons; control task versus L-counting: F(1, 11) = 10.0; CMT versus L-counting: F(1, 11) = 6.1, ps < .05. Steering intervals in the control condition and the CMT did not differ, F(1, 11) = 1.0, p > .20. In the loading task period, the average steering interval across all loading task conditions amounted to 3.6 s, which decreased to 3.0 s when the secondary task was also carried out, F(1, 11) = 7.2, p < .05. This indicates that the 3.3-s steering interval in the L-counting condition was not the lowest value possible. Hence, steering intervals were not subject to a floor effect and seem useful for indicating still higher levels of visual workload.

Workload Assessment Techniques

Secondary workload performance. Figure 3 and Table 1 show that secondary task performance dropped when a loading task was carried out, whereas performance remained unchanged in the control condition. Planned comparisons showed that this performance drop was significant for the L-counting and the CMT conditions separately, F(1, 11) = 28.0, p < .001, and F(1, 11) = 12.3, p < .01, respectively. Moreover, the difference between L-counting and the CMT was significant, F(1, 11) = 25.9, p < .001.

Figure 4 demonstrates that the percentage of detected targets decreased as L-counting and CMT
DETECTING SHORT WORKLOAD PERIODS

Figure 3. Secondary task performance as a function of period and loading task. CMT = continuous memory task; L = L-shaped light-emitting diodes.

lasted longer, but this was not the case in the control condition, \( F(1, 11) = 6.8, p < .05 \). Planned comparisons per duration confirmed this, as secondary task performance was lower at 60-s L-counting than at 10- and 30-s L-counting, \( F(1, 11) = 6.6 \) and \( F(1, 11) = 8.9 \), respectively, \( p < .05 \), and secondary task performance was higher at 10-s CMT than at 30- and 60-s CMT, \( F(1, 11) = 7.3, p < .05 \), and \( F(1, 11) = 12.4 \), respectively, \( p < .01 \). It appears that timesharing a loading task and the secondary task became more difficult, as this situation took longer.

Subjective workload. As shown in Table 2, SWAT and RSME were both affected by loading task and duration. Furthermore, Loading Task \( \times \) Duration interactions were found for both subjective measures, \( F(4, 44) > 4.0, p < .01 \), suggesting that the duration effect occurred with L-counting and CMT, but not in the control condition (see Figure 5). Pairwise comparisons of the three loading task conditions showed that SWAT and RSME were both sensitive to differences between control condition and L-counting and between control condition and CMT (see Table 2).

Table 2 also shows that both subjective workload measures produced virtually the same results with respect to duration. However, there was a difference in that SWAT differentiated between 10- and 30-s L-counting, which difference was not significant for RSME.

IBIs and HR variability. As indicated in Table 1, IBI and HR variability were not very sensitive to workload. Planned comparisons showed a marginally significant tendency of IBI to reduce when L-counting and CMT took longer, whereas this was not the case in the control condition, \( F(1, 11) = 4.2, p = .06 \) (see Figure 6). With 10-s L-counting, IBIs were longer than with 60-s L-counting, \( F(1, 11) = 6.6, p < .05 \). During the 60-s CMT

Table 2
Statistical Significance of the Differences Between Conditions as Indicated by the Subjective Rating Scales SWAT and RSME

<table>
<thead>
<tr>
<th>Condition</th>
<th>SWAT</th>
<th>RSME</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loading task (main)</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>No-loading task vs. L-counting</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>L-counting vs. CMT</td>
<td>&lt;.10</td>
<td>&gt;.10</td>
</tr>
<tr>
<td>No-loading task vs. continuous memory task</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Duration (main)</td>
<td>&lt;.001</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>No-loading only*</td>
<td>&gt;.10</td>
<td>&gt;.10</td>
</tr>
<tr>
<td>L-counting only (s)</td>
<td>&lt;.001</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>10 vs. 30P*</td>
<td>&lt;.001</td>
<td>&gt;.10</td>
</tr>
<tr>
<td>10 vs. 60</td>
<td>&lt;.001</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>30 vs. 60</td>
<td>&gt;.10</td>
<td>&gt;.10</td>
</tr>
<tr>
<td>Continuous memory task only (s)</td>
<td>&lt;.001</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>10 vs. 30</td>
<td>&lt;.05</td>
<td>&lt;.10</td>
</tr>
<tr>
<td>10 vs. 60</td>
<td>&lt;.001</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>30 vs. 60</td>
<td>&gt;.10</td>
<td>&gt;.10</td>
</tr>
</tbody>
</table>

Note. SWAT = Subjective Workload Assessment Technique; RSME = Rating Scale Mental Effort; L-counting = number of L-shaped light-emitting diodes.

*Duration effects per loading task tested with 10 versus 60-s planned comparisons. * Differences between durations tested with Tukey post hoc tests.
Figure 5. Subjective ratings as a function of loading task and duration. L = L-shaped light-emitting diodes; RSME = Rating Scale Mental Effort; SWAT = Subjective Workload Assessment Technique.

Figure 6. Effects of loading task and duration on interbeat intervals of the heart beat. CMT = continuous memory task; L = L-shaped light-emitting diodes.

periods, IBIs were shorter than with 10- and 30-s CMT, $F(1, 11) = 19.8$, and $F(1, 11) = 19.0$, $p < .001$, respectively.

HR variability was lower in 60-s L-counting periods than in 60-s control condition periods; 0.049 versus 0.059, $F(1, 11) = 8.1$, $p < .05$. Likewise, HR variability was lower when the CMT was performed for 60 s as compared with 60-s control condition periods; 0.050 versus 0.059, $F(1, 11) = 5.6$, $p < .05$. Finally, HR variability was lower in the no-loading task periods of driving than while standing still; 0.057 versus 0.084; $F(2, 22) = 9.3$, $p < .05$.

Eyeblinks. L-counting was not included in the eyeblink analysis because eyeblinks could not be distinguished from the rapid vertical eye movements made toward the L-counting display. The average eyeblink intervals in the CMT and the control condition were not different; 6.7 versus 7.0 s; $F(1, 11) = 0.2$, $p > .20$. However, loading task period duration had different effects for the CMT and the no-loading task: in the no-loading task condition eyeblink intervals decreased with duration from 7.6 (10 s) to 7.3 (30 s) to 7.0 s (60 s), whereas in the CMT conditions, eyeblink intervals increased from 6.4 s to 7.0 to 8.2 s, $F(1, 11) = 7.0$, $p < .05$.

SCRs. Table 1 suggests that SCR responded to the loading tasks. Planned comparisons showed that the intervals between successive SCRs were shorter in the CMT condition than in the control and L-counting conditions; CMT versus control: $F(1, 11) = 18.9$, $p < .001$; CMT versus L-counting, $F(1, 11) = 8.8$, $p < .05$ (see Figure 7). Intervals between SCRs in the L-counting and control conditions were not different, $F(1, 11) = 2.4$, $p > .10$. There was a marginally significant difference between SCR intervals while standing still and when driving in the no-loading task period; 24.2 s versus 16.2 s; $F(2, 22) = 3.1$, $p = .06$.

Comparing Workload Measures

Correlations and factor analyses. Table 3 shows correlations between the various workload mea-
Figure 7. Effects of loading task and period on skin conductance responses (SCRs). CMT = continuous memory task; L = L-shaped light-emitting diodes.

Table 3

Correlations Between the Various Measures for the Control and L-Counting Conditions and the Control and Continuous Memory Task (CMT) Conditions

<table>
<thead>
<tr>
<th>Measure</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. SWAT</td>
<td>—</td>
<td>.88</td>
<td>.61</td>
<td>.70</td>
<td>.69</td>
<td>-.37</td>
<td>-.02</td>
<td>-.60</td>
<td>-.20</td>
<td>.07</td>
<td>-.08</td>
</tr>
<tr>
<td>2. Time</td>
<td>.77</td>
<td>—</td>
<td>.36</td>
<td>.49</td>
<td>.63</td>
<td>-.37</td>
<td>.05</td>
<td>-.55</td>
<td>-.16</td>
<td>.04</td>
<td>-.10</td>
</tr>
<tr>
<td>3. Effort</td>
<td>.85</td>
<td>.59</td>
<td>—</td>
<td>.05</td>
<td>.50</td>
<td>-.18</td>
<td>-.19</td>
<td>-.43</td>
<td>-.04</td>
<td>.18</td>
<td>.04</td>
</tr>
<tr>
<td>4. Stress</td>
<td>.66</td>
<td>.17</td>
<td>.34</td>
<td>—</td>
<td>.37</td>
<td>-.23</td>
<td>.07</td>
<td>-.33</td>
<td>-.23</td>
<td>-.05</td>
<td>-.10</td>
</tr>
<tr>
<td>5. RSME</td>
<td>.66</td>
<td>.53</td>
<td>.66</td>
<td>.30</td>
<td>—</td>
<td>-.28</td>
<td>.04</td>
<td>-.62</td>
<td>-.21</td>
<td>.04</td>
<td>-.15</td>
</tr>
<tr>
<td>6. Steer</td>
<td>-.26</td>
<td>-.35</td>
<td>-.21</td>
<td>-.04</td>
<td>-.23</td>
<td>—</td>
<td>-.13</td>
<td>.33</td>
<td>.03</td>
<td>-.13</td>
<td>.06</td>
</tr>
<tr>
<td>7. SCR</td>
<td>-.05</td>
<td>-.07</td>
<td>-.03</td>
<td>-.01</td>
<td>-.14</td>
<td>-.19</td>
<td>—</td>
<td>.07</td>
<td>-.07</td>
<td>.14</td>
<td>-.25</td>
</tr>
<tr>
<td>8. Secondary task</td>
<td>—</td>
<td>-.44</td>
<td>-.20</td>
<td>-.53</td>
<td>-.28</td>
<td>-.44</td>
<td>.10</td>
<td>.08</td>
<td>—</td>
<td>.03</td>
<td>-.03</td>
</tr>
<tr>
<td>9. Speed</td>
<td>-.03</td>
<td>-.03</td>
<td>-.07</td>
<td>-.04</td>
<td>-.02</td>
<td>-.05</td>
<td>.03</td>
<td>.16</td>
<td>—</td>
<td>-.36</td>
<td>.15</td>
</tr>
<tr>
<td>10. Interbeat interval</td>
<td>—</td>
<td>-.04</td>
<td>-.02</td>
<td>.13</td>
<td>-.20</td>
<td>-.07</td>
<td>-.22</td>
<td>.20</td>
<td>.04</td>
<td>-.43</td>
<td>—</td>
</tr>
<tr>
<td>11. Heart rate variability</td>
<td>.02</td>
<td>.12</td>
<td>-.02</td>
<td>-.06</td>
<td>-.09</td>
<td>.07</td>
<td>-.29</td>
<td>.36</td>
<td>.15</td>
<td>-.07</td>
<td>—</td>
</tr>
<tr>
<td>12. Eye blink rate</td>
<td>.19</td>
<td>-.11</td>
<td>.12</td>
<td>.43</td>
<td>.10</td>
<td>.18</td>
<td>.34</td>
<td>-.18</td>
<td>-.08</td>
<td>-.09</td>
<td>-.30</td>
</tr>
</tbody>
</table>

Note. Each correlation is based on the average workload index per participant in the control condition and in the CMT or L-counting (number of L-shaped light-emitting diodes) loading task condition for each duration (i.e., \( N = 12 \times 2 \times 3 = 72 \) observations); significant correlations are boldface; blink rate is not indicated for L-counting because rapid vertical eye movements associated with scanning the L-counting display could not be distinguished from eyeblinks. L-counting conditions are in the upper right and CMT conditions are in the lower left. SWAT = Subjective Workload Assessment Technique; RSME = Rating Scale Mental Effort; SCR = skin conductance response.

*No comparable data in the L-counting condition.
ing intervals, and SCR intervals were compared in a Technique × Loading Task × Duration ANOVA after all data had been z-transformed. Figure 8 depicts the z scores for these promising workload measures. Table 4 indicates statistical significance of the differences in sensitivity between techniques. The secondary task was more sensitive to 10-s L-counting periods than any other task. Overall, SCR was a worse indicator for the workload of 60-s L-counting periods than SWAT, RSME, and the secondary task. Workload induced by executing the CMT for 10 s was indicated poorly by all measures. CMT performance of 30 and 60 s was indicated better by RSME than by steering intervals, SCR, and secondary task performance.

There are indications that workload might also have an effect on periods after workload elevations have ceased (Hancock, Williams, Manning, & Miyake, 1995; see also our Figure 1). This hypothesis was tested by comparing measures obtained in the no-loading task periods of the control condition, L-counting and CMT. This analysis was carried out for speed, steering intervals, secondary task performance, IBI, HR variability, and SCR intervals. It yielded a marginally significant effect for steering intervals only, $F(1, 11) = 4.4, p = .06$ (see Figure 2). Hence, the present study provides little support for the notion that relatively short workload periods have lasting effects.

Finally, Table 5 was produced to depict which measures are useful for finding differences at each workload duration. This table shows that secondary task performance, SWAT, and RSME were sensitive to the workload imposed by both the L-counting task and the CMT at all three durations with the exception that the SWAT could not distinguish 10-s CMT from the control condition. Table 5 also shows that HR variability differentiated significantly between driving and standing still. SCR showed a marginally significant difference. The IBI did not indicate a difference between driving and remaining still.

**Discussion**

The present study investigated the sensitivity and the diagnosticity of nine common workload assessment techniques used to evaluate shorter and longer periods of elevated visual and mental workload. Sensitivity refers to the extent that a technique discriminates between different levels of workload. Diagnosticity indicates whether a technique responds differently to different types of workload.

The results were clearcut. Secondary task performance, SWAT, and RSME were sensitive to the visual 10-s workload peaks caused by L-counting. Secondary task performance, RSME, and, somewhat less reliably, SCR were sensitive to the 10-s mental workload peaks induced by the CMT. In addition, steering intervals and HR variability were sensitive to longer (30 and 60 s) periods of L-counting, whereas SWAT and HR variability were sensitive to longer expositions to the CMT.

The secondary task and steering intervals were more sensitive to L-counting than to the CMT; therefore, these techniques are diagnostic with respect to the type of workload. However, the data also showed that the performance reduction of the visually loading secondary task does not necessar-

![Figure 8](image-url)
DETECTING SHORT WORKLOAD PERIODS

Table 4
Comparisons of the Sensitivity of Pairs of Measures as a Function of Loading Task and Task Duration

<table>
<thead>
<tr>
<th>Measure</th>
<th>L-counting* (p)</th>
<th>CMT (p)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10 s</td>
<td>30 s</td>
</tr>
<tr>
<td>SWAT vs. RSME</td>
<td>&gt;.10</td>
<td>&gt;.10</td>
</tr>
<tr>
<td>SWAT vs. steering</td>
<td>&gt;.10</td>
<td>&gt;.10</td>
</tr>
<tr>
<td>SWAT vs. SCR</td>
<td>&gt;.10</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>SWAT vs. secondary task</td>
<td>&lt;.01</td>
<td>&gt;.10</td>
</tr>
<tr>
<td>RSME vs. steering</td>
<td>&gt;.10</td>
<td>&gt;.10</td>
</tr>
<tr>
<td>RSME vs. SCR</td>
<td>&gt;.10</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>RSEM vs. secondary task</td>
<td>&lt;.05</td>
<td>&gt;.10</td>
</tr>
<tr>
<td>Steering vs. SCR</td>
<td>&gt;.10</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>Steering vs. secondary task</td>
<td>&lt;.01</td>
<td>&gt;.10</td>
</tr>
<tr>
<td>SCR vs. secondary task</td>
<td>&lt;.001</td>
<td>&lt;.01</td>
</tr>
</tbody>
</table>

Note. Levels of significance are as indicated by Technique x Loading Task (L-counting or CMT vs. control) interactions in separate analyses of variance. CMT = continuous memory task; SWAT = Subjective Workload Assessment Technique; RSME = Rating Scale Mental Effect; SCR = skin conductance response. *Number of times light-emitting diodes formed an L shape.

...ily point to increased levels of visual workload alone because secondary task performance was also affected by the CMT. SCR responded only to mental workload, but its sensitivity was limited, given the fact that the statistical significance of these effects was "at the edge." In line with earlier results (O'Donnell & Eggemeier, 1986), the subjective techniques SWAT and RSME appeared not to be diagnostic with respect to visual and mental workload in that they were about equally sensitive to L-counting and CMT. The observation that some measures were more sensitive to visual or mental workload underlines that workload is a multidimensional concept (Wickens, 1984b) and that, dependent on the question at hand, the diagnosticity of a workload assessment technique should be taken into account when designing a workload study.

We found that HR variability distinguished between driving with and without a 60-s loading...
task, but it appeared not to be sensitive to 10- and
30-s periods of elevated workload. Moreover, HR
variability discriminated between driving without a
loading task and remaining still. This suggests that
HR variability is sensitive to longer periods of
elevated workload, but that this measure is not
suitable for detecting workload elevations of less
than 1 min.

As noted in the Method section, the SWAT was
analyzed in four different ways. These analyses
yielded virtually identical results; hence, the re-
sults of only one type of analysis were reported in
subsequent analyses. Still, it is important to realize
that even the simplest analysis, adding the three
ratings, did about as well as the analyses that
included individual differences, as obtained by the
card-sorting procedure. This confirms the results
of researchers who worked with homogeneous
participant groups (see Moroney, Biers, & Egg-
gemeier, 1995, for an overview), for the heteroge-
neous group of drivers in the present study. More-
ever, the present data showed that the RSME
could reliably detect the workload associated with
10-s CMT periods. This was not the case with the
SWAT. In other words, it seems that in driving
studies, subjective workload assessment tech-
niques gain little from complex assessment and
analysis methods such as those used in the SWAT.
A simple one-dimensional rating scale appears
sufficient for assessing subjective workload in driv-
ing. Future research should investigate whether
there are specific task environments where com-
plex scales like the SWAT and the National
Aeronautics and Space Administration Task-Load
Index (NASA-TLX; Hart & Staveland, 1988) are
beneficial.

The highly limited sensitivity of eyeblinks to the
CMT, together with findings in other studies that
eyeblinks were sensitive to visual workload (e.g.,
Veltman & Gaillard, 1996; Wilson & Fisher, 1990),
corresponds with earlier claims that eyeblinks are
diagnostic with respect to the type of workload
(Casali & Wierwille, 1984; Wierwille & Eggemeier,
1993). The EOG could not distinguish between
rapid vertical eye movements toward the in-vehicle
display and eyeblinks. Therefore, if drivers repeat-
eedly glance at in-car displays, applied researchers
might be better off with observation techniques to
determine eyeblink frequency than with EOG.

An important theoretical result is that the factor
analysis revealed two independent factors to under-
lie visual and mental workload. The first factor
involved subjective workload ratings (SWAT, RSME,
and the time pressure dimension of the
SWAT) and, in case of L-counting, secondary task
performance. The second factor included the physi-
ological measures SCR, HR variability, and, in the
CMT condition, eyeblink rate. This supports the
suggestion in the introduction that subjective rat-
ings and secondary task performance index inform-
ation-processing load, or task complexity, and
physiological measures indicate the effort invested
in the task or, perhaps also, the stress caused by
attending to a task unrelated to driving (cf. Casali

To help readers determine the most appropriate
workload assessment technique for studies on
visual and mental workload, we present in Table 6
a simplified and pragmatic overview of the pres-
ently used workload assessment techniques in
terms of their sensitivity and diagnosticity.

It is important to note that the conclusions with
respect to the secondary task should be considered
with caution because intrusion of driving by the
secondary task was measured only with respect to
speed and steering frequency. Although speed did
not show an effect of secondary task performance,
steering frequency did increase. The data do not
indicate the extent to which other parts of the
driving task, such as braking reaction time and lane
keeping, were affected. Various researchers (e.g.,
Noy, 1987) have argued that secondary tasks are
likely to affect primary task performance. This
stresses that participants should always and repeat-
eedly be instructed to stop performing the second-
ary task whenever driving performance tends to
deteriorate (Wickens, 1984a). The use of rewards
and immediate performance feedback may be of
value here. Wierwille and Gutmann (1978) found
that primary task performance was affected more
by a secondary task, as the primary task was less
demanding. Therefore, when the secondary task

A second limitation of the present study is that
the loading tasks involved only one level of diffi-
culty (i.e., absence vs. presence). This makes it
difficult to estimate the sensitivity of the present
Table 6
Sensitivity and Diagnosticity of Visual and Mental Workload Measures Used in This Study and the Durations for Which These Qualifications Hold

<table>
<thead>
<tr>
<th>Measure</th>
<th>Visual</th>
<th>Mental</th>
<th>Diagnosticity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steering frequency</td>
<td>High $\geq 30,\text{s}$</td>
<td>Poor</td>
<td>High $\geq 30,\text{s}$</td>
</tr>
<tr>
<td>Secondary task(^a)</td>
<td>High $\geq 10,\text{s}$</td>
<td>High $\geq 10,\text{s}$</td>
<td>Reasonable $\geq 10,\text{s}$</td>
</tr>
<tr>
<td>SWAT</td>
<td>High $\geq 10,\text{s}$</td>
<td>High $\geq 30,\text{s}$</td>
<td>None</td>
</tr>
<tr>
<td>RSME</td>
<td>High $\geq 10,\text{s}$</td>
<td>High $\geq 10,\text{s}$</td>
<td>None</td>
</tr>
<tr>
<td>Skin conductance response</td>
<td>None</td>
<td>Reasonable $\geq 10,\text{s}$</td>
<td>Reasonable $\geq 10,\text{s}$</td>
</tr>
<tr>
<td>Heart rate variability</td>
<td>Reasonable $\geq 60,\text{s}$</td>
<td>Reasonable $\geq 60,\text{s}$</td>
<td>None</td>
</tr>
<tr>
<td>Interbeat interval</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Speed</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Eyeblinks</td>
<td>Yes(^b)</td>
<td>None</td>
<td>Reasonable/High(^b)</td>
</tr>
</tbody>
</table>

Note. SWAT = Subjective Workload Assessment Technique; RSME = Rating Scale Mental Effort.

\(^a\)Qualifications depend on type of secondary task. \(^b\)Expected on basis of other studies (e.g., Veltman & Gaillard, 1996; Wilson & Fisher, 1990).

workload assessment techniques to different levels of workload. In the extreme, the present techniques may have responded to the presence of the loading tasks per se, whereas the actual difficulty level was unimportant. Future research should assess the sensitivity of the techniques in this study to more than the two levels of workload reported in this article.

An important characteristic of the present study design was that the moments of workload elevation were known to the experimenter. However, workload assessment techniques will be less sensitive when the moments of elevated workload are not recognized and measures are assessed over much longer periods of time. Then, the occurrence of workload peaks may be entirely averaged out (Knowles, 1963; Wierwille, 1981). Determining whether workload peaks occur calls for application of these three methods: (a) executing a task analysis prior to the experiment and assuring that workload is captured at the appropriate moments; (b) in case of continuous techniques (secondary task performance and physiological measures) computing moving averages (Antin & Wierwille, 1984; Wierwille, 1981), and (c) asking participants to indicate workload peaks and their levels (Wierwille & Eggemeier, 1993).

With respect to use of subjective rating scales, it is important to note that these techniques usually do not instruct participants whether they should estimate peaks or averages across longer periods of time. The consequence of this omission is an increase of individual differences because different participants will understand the rating task differently. Hence, the researcher should always indicate explicitly whether workload ratings should regard peaks or averages and to which period of time or to which situation the estimate should refer.

Another theoretical issue concerns the relationship between workload and performance. Tasks may differ considerably with respect to the probability that performance is affected by peak workload. For example, driving always involves some time pressure, and the occurrence of peak workload is probably more important for task performance than for tasks without time pressure such as in human-computer interactions. This suggests that researchers should adopt a workload assessment technique that is more sensitive to peak workload when task performance is more likely to be affected by peak workload. For driving research, assessment of peak workload seems crucial. Secondary task performance and subjective ratings appear to be appropriate techniques in this respect because they are sensitive to workload peaks.

In conclusion, the present study demonstrates that subjective ratings and secondary task techniques are good methods for assessing visual and mental workload peaks in driving. The relatively simple univariate RSME rating scale was somewhat more sensitive than the more complex SWAT. Steering frequency appears to be a sensitive measure for longer periods of visual workload, and
SCR appears to be sensitive to mental workload. The present results suggest that physiological measures such as SCR, HR variability, and eyeblink rate reflect invested effort rather than information-processing load and task performance. Because driving is a task in which peak workload may have serious ramifications for safety, studies on workload in driving should adopt workload assessment techniques that can evaluate the occurrence of peak workload. The present results show that secondary task performance and subjective ratings are useful in this respect. In task environments with less time pressure, this requirement may be less important.

References


DETECTING SHORT WORKLOAD PERIODS

cock & N. Meshkati (Eds.), Human mental workload (pp. 185–218). Amsterdam: North Holland.

Received May 3, 1995
Revision received February 26, 1996
Accepted March 1, 1996