

## Conclusions related to driver workload measures

---

In the previous chapter the characteristics of different workload measures in traffic research, and in particular in car driving, were evaluated. The most important characteristic of a measure is its sensitivity to workload. Workload and task demand were explicitly separated, the former reflecting the individual reaction to the latter. It was argued that the sensitivity of a measure is highly dependent upon region of performance. Outside the regions in which a measure is sensitive, ceiling or floor-effects occur, which may give the impression that a measure is insensitive overall. Some researchers actually have (mis)interpreted these effects as dissociation of measures. Apart from sensitivity, reliability and diagnosticity are important characteristics of measures. A highly diagnostic measure is selectively sensitive, e.g. to visual workload, or will indicate at what processing stage mental workload is increased. In particular, diagnosticity has strong links to the multiple-resource theory. In studying the effects on workload of navigation aids, a diagnostic measure can be required. Measures can be highly sensitive to, for instance, visual information processing in the encoding stage only, which can be important if the effects of a visual display are studied. The measure may not respond to increases in workload at other stages, and should be insensitive to workload in, for instance, the response-choice stage. Hence, diagnosticity restricts sensitivity to a certain bandwidth. Global workload measures are low in diagnosticity and provide few clues about the stages in which demand for resources are high. On the other hand, these measures are useful in the assessment of overall workload.

Of primary importance in mental workload research is the region of performance (figure 2). Optimal performance with low mental workload is obtained in region A2. If the driver's state is affected, e.g. after the use of sedative drugs, the driver might (at first) successfully counteract these negative consequences by the investment of (state-related) effort. The performance level on the primary task remains unaffected, but in particular self-report ratings on the RSME may indicate increased costs, while the 0.10 Hz component of heart rate variability does not seem to be significantly sensitive to state-related effort. Performance is said to be in region A1. If effort compensation is no longer possible, performance will deteriorate which will be reflected by the primary task measures. EEG and ratings on the activation scale also mirror an affected state. From the ECG measures the average heart rate level seems to be most sensitive, but mainly to effects of time-on-task.

Not only driver state, but also the complexity of a task or of the driving environment may cause an increase in mental workload. Again, optimal performance is in region A2. If the traffic environment

becomes more complex, e.g. when a weaving section is passed, or if a secondary task is added to the driving task, e.g. messages from a Feedback device, then drivers have to exert (task-related) effort to maintain performance. Both the 0.10 Hz component of heart rate variability and the self-report scale RSME seem to be able to indicate task related effort. In the previous chapter it was also suggested that as a 'side-effect' of this effort compensation, performance on the primary task might even increase. Once the driver is no longer able to successfully act against detrimental effects of increased task demands by means of effort compensation, the level of performance will drop and performance is said to be in Region B. In particular performance measures will indicate this. With further increases in demand, performance will further drop until a minimum level is reached and Region C is entered. In this region, performance measures are insensitive, nor will measures of heart rate reflect workload. Only scores on the activation scale may indicate overload.

What is clear from the above is that none of the measures alone is sufficient to reflect mental workload. An identical performance level may indicate optimal performance, effort compensation or overload. Only in combination with self-reports and/or physiological parameters can a conclusion about workload level be made. Very important in mental workload assessment are individual differences and strategies. Even if task demands are equal for two persons, their reaction to the demands -how difficult the task at hand is for them- may very well differ. This complicates generalisation, as in principle, a task cannot be assigned to a region of demand (figure 2) in advance. Individual goal-setting and the interaction between complexity and capability, and thus difficulty, differ between individuals.

In recent years, the question 'How much workload is too much' has received increased attention. In an applied setting such as traffic research, the workload redline could be a very useful concept as the consequences of too much workload in driving can be very serious. In this thesis I have questioned the correctness of putting the redline at the point at which performance is affected and have suggested as alternative the point of time at which effort compensatory processes are initiated. For this, the combination of performance measures with physiology and/or self-report measures can provide a picture of mental workload. Critical levels of measures of mental workload are, however, not attainable as mental workload itself is a relative measure. The resources the operator is willing or capable to allocate to task performance differ between individuals and make a redline in the form of a critical level on a measure of mental workload impossible. Changes in strategy and the self-pacedness of the driving task add to this. For example, the SDLP performance measure and self-report measures may remain unaffected under conditions of increased task demands simply because the driver has adapted task difficulty by

driving at a slower speed. This does not mean that performance 'as a whole' remains unaffected as one of the performance measures, speed, should reflect this change in strategy.

While critical levels of mental workload are difficult to determine, absolute critical levels of performance that are considered thresholds for unaffected performance can be determined because performance is an objective measure. These measures are not workload redlines, but primary-task workload margins (Wickens, 1984). Although this approach is more likely to be successful than workload redline determination, it should be stressed again that unaffected performance is not equal to low mental workload. Prolonged effort compensation may exclude effects on performance measures, but could be a threat to good health. It has, for instance, been suggested that repetitive activation of the cardiovascular defense response (i.e., task-related effort) may lead to hypertension (see Johnson & Anderson, 1990).

Most of the primary-task performance measures have been strongly linked to control-level processes. Control-level processes on their part have been linked to automatic processes and these processes are said to require hardly any resources. A reduction in capacity, e.g. as a result of the use of alcohol, should leave most automatic processes unaffected. In fact many control level aspects of driving remain unaffected after consumption of low amounts of alcohol. In this respect, the sensitivity of the primary task measures SDLP and SDSTW to, for instance, low amounts of alcohol is unexpected. However, it is of primary importance to acknowledge that most tasks have both automatic and controlled aspects. Or as Schneider and Fisk (1983) stated: 'there is rarely any task in which processing is purely controlled or purely automatic'. The sensitivity of the primary-task measures could be the result of the controlled processing component, e.g., the degree to which the driver cuts corners. The automatic processes are the process-components that execute the appropriate movements, i.e. the steering actions.

It should be noted that mental load was used in a broad context in the different studies. The results of a variety of experiments were used for this evaluation, and evaluation of the mental load measurement technique itself was not the original research issue in these studies. All studies were field tests or studies in which the driving environment was simulated, and subjects were always seated in a real vehicle. Techniques that in most cases had been developed in the laboratory were tested in an applied environment; out on the road. The results of this transition sometimes showed that measures were very sensitive in the field. For example, the 0.10 Hz component of heart rate variability reflected remarkably well changes in road environment and task complexity in the Weaving Section and simulator study. In these studies, and others (car-phone study, the Woodland Road in the road-

layout experiment), task-task differences, i.e. differences between baseline condition and load condition, were found. Jorna's conclusion (Jorna, 1992) that the 0.10 Hz component of HRV is only sensitive to large task differences is therefore not supported. In particular, the profile technique can be very useful in the evaluation of ongoing changes in mental effort. I therefore disagree with Grossman (1992) who considers the measure 'interesting' but questions its validity due to lack of understanding of the complete underlying physiological basis. As long as it is not exactly understood what the variable represents, he considers use of the 0.10 Hz component doubtful. It is, however, by no means true that a measure cannot be useful until the complete (physiological) mechanisms are understood. For instance, we do not understand how a subject introspects and rates the amount of effort invested, yet self-report measures have proven to be very useful in workload research. Moreover, a plausible explanation for the 0.10 Hz HRV-rhythm in terms of a relation to a decreased baroreceptor reflex sensitivity has been offered (see G.Mulder, 1980, L.J.M.Mulder, 1988).

### **Region of performance and measure's sensitivity to workload**

We have used measures from the three measurement categories: task-performance measures, self-report measures and physiological measures. Which technique to choose should depend primarily on the research question. That is to say, it *should* depend upon the research question, although in practice, the researcher conducting a field experiment will find him or herself limited by many constraints. Specialised equipment for the measurement of physiological signals and expensive instrumented cars are required for the assessment of changes in CNS activity and primary-task performance respectively. The need for this specialized equipment has made the use of self-report questionnaires very popular. Again, it should be stressed that on their own these reports can indicate mental workload only to a restricted degree. To obtain a complete picture, and to be able to assess region of performance, measures from at least one other category are required. In research, the use of a test battery, or a minimum of more than one measurement technique, is therefore advised. More than that, in complex environments, it is advisable to use more measures from the same measurement category. To quote Wilson & Eggemeier (1991), "It seems that the strategy of recording only a single physiological variable, such as heart rate, is no longer appropriate in most multi-task studies". Meijman & O'Hanlon (1984) state: "Just as there are multiple causes of mental workload, there are multiple effects". Their advice to the applied scientist is to identify and control as many sources of mental workload as possible, and to measure performance, physiology, and gather self-report ratings simultaneously.

Sometimes, general statements about measurement techniques of different categories regarding their region-sensitivity are made, e.g.,

primary-task performance insensitivity in the A region (O'Donnell & Eggemeier, 1986). However, measures have differential sensitivities, even within the same category. For instance, two ECG measures, heart

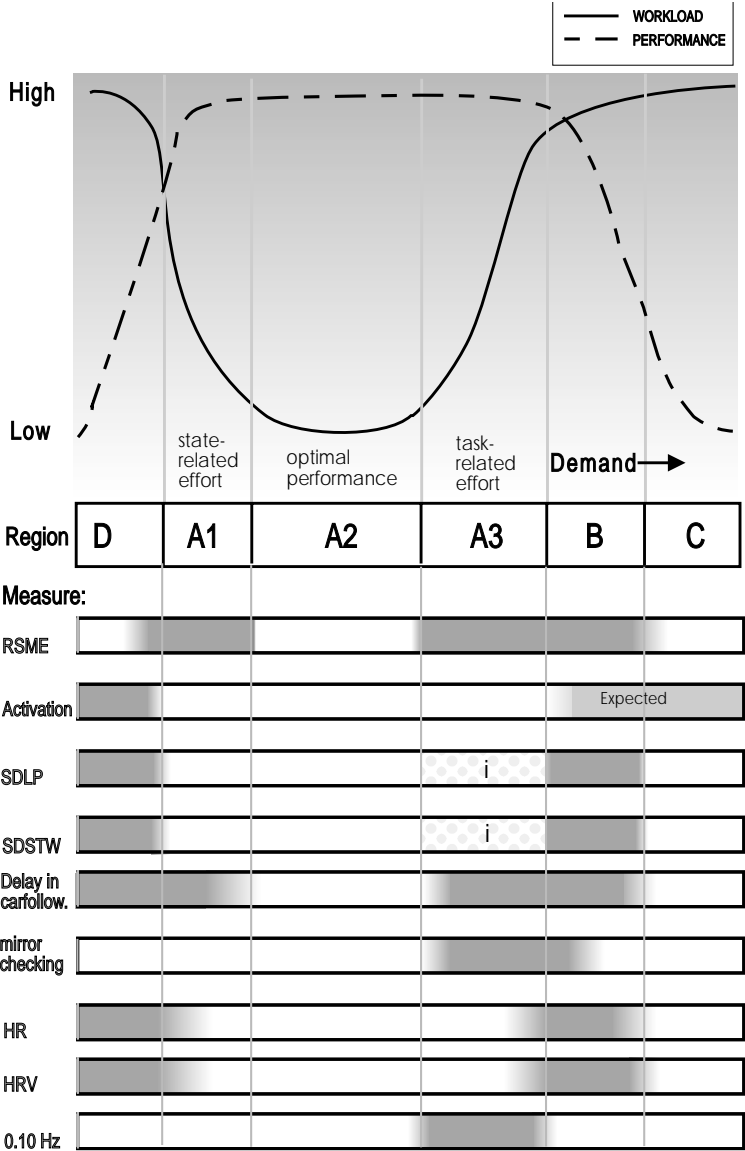


Figure 19. Workload in 6 regions and sensitivity of different measures to driver mental workload. RSME and Activation are self-report scales, SDLP and SDSTW are primary-task performance parameters. 'Delay carfollow.' is an embedded secondary task measure; the delay in following of speed changes of a lead car. 'Mirror checking' reflects sensitivity of the embedded measure 'frequency of mirror checking', while HR, HRV and .10 Hz are cardiac measures. The meaning of 'i' (improvement possible) is discussed in the text.

rate (HR) and the 0.10 Hz component of heart-rate variability, are both physiological measures. The 0.10 Hz component of heart rate variability is most sensitive in the A3-region, while HR itself is more sensitive in the D and B regions. Likewise, self-report activation scales may be sensitive in the D region, while mental effort scales may not be sensitive there. For this reason, in figure 19 the measures we have used most frequently are listed individually.

Driving is to a large extent a self-paced task. The implication of this is that the task in terms of performance achieved is varied (Parkes, 1991). However, some of the paced aspects can be captured, e.g. in terms of average driving speed chosen. Some conclusions with respect to compensatory behaviour can be based on these observations. The self-pacedness of the driving task could also account for the improvement in performance that was found in the car-phone study, road layout and DETER studies on one of the primary-task measures, the SDLP, in the load conditions. Gopher and Sanders (1984) have suggested for unexpected primary task performance improvement a similar explanation in terms of a change in task emphasis (in dual task performance), or resource allocation policy.

### **Recommendations for driver mental load measurement**

- *Multiple measures.* Measures from different categories should be used. If possible, this should include multiple measures within categories.
- *Self-report measures.* To reduce primary-task interference, questionnaires and scales should be filled in after completion of the task. Depending upon the research question, a decision regarding the use of a multidimensional or unidimensional scale should be made: if overall workload has to be assessed, the unidimensional scale is to be preferred. If driver state can be affected, an activation rating is useful in addition to ratings of workload or effort.
- *Primary-task performance measures.* Primary-task performance measures are very important for mental load assessment for two reasons. Firstly, reduced primary task performance can indicate overload or a reduced driver state. Secondly, improved performance could be the result of a change in task interpretation and/or effort-compensatory processes.

Primary-task performance measures have to be carefully selected, and all suffer from problems. In the field, general measures such as time-to-complete a circuit are susceptible to many disturbing factors. Steering-wheel measures can only be applied at specific locations, i.e. at spots where the road curvature is known or nil. Finally, the SDLP and TLC-measures can only be applied on roads with a delineation and both require specialized equipment (e.g., a 'lane

tracker'). If the SDLP measure is used, care has to be taken that driving speed between conditions is comparable (e.g., Godthelp, 1988) and that roads have the same lane width (e.g., Green et al., 1993b).

Driving speed can reflect changes in goal setting. A slower driving speed may be an individual adaptation to be better able to deal with the task demands and this slower driving speed may 'mask' effects on other parameters. Registration of driving speed is thus important but mainly in the function of control parameter registration.

- *Secondary-task performance measures.* Added tasks have as major disadvantage that they interact with primary-task performance. The best secondary tasks to use in the field are embedded secondary-task measures, such as frequency of mirror-checking and car-following performance.

- *Physiological measures.* Measurement of physiological signals necessitates some expertise and specialized equipment. Heart rate measurements have been applied in the field for some time now and new techniques such as the profile technique offer the possibility of monitoring changes in workload during performance. If physiological measures are taken then rest measurements have to be included for scaling and to assess resting baseline physiological activity. In particular in a test-environment these resting-baselines can be affected, especially in highly anxious or reactive subjects, making interpretation and comparisons between studies only meaningful if rest measurements are included (Papillo & Shapiro, 1990). The Law of Initial Values also states that the range of responses will be restricted in case of a high resting baseline. A combination of before and after-test resting-baselines may help to restrict this effect. Ultimately, however, in driver mental load assessment the measurements gathered in a baseline *driving* condition should be used to compare measures in a load condition with. Scaling of these two measurements should be based on the rest-measurements. A simple way to do this for power spectral density analysis performed on heart rate data, is to logarithmically transform the spectral values (Van Roon, in preparation). This transformation also leads to a normal distribution of data.

It is advised that the driver remains silent while driving when ECG measures are taken (e.g., avoid verbal ratings), although there are reports that a limited number of vocalizations do not disturb heart rate measures (Porges & Byrne, 1992).

Facial EMG may be a promising measure in the domain of mental workload, but very few studies that included the measure have been performed in the field. EEG is very useful to assess driver state.

- *Experimental Design.* In setting up a field experiment in which mental load has to be assessed, inclusion of the following aspects should be considered;

- Straight road segments for steering wheel measures.
- Comparable (preferably identical) baseline and load conditions in terms of selected test-road and traffic density.
- For heart rate and other physiological measures: before-task (or between tasks), and after-task rest-measurements.

### **Applied research**

One of the disadvantages of field experiments is that there is no control over what happens in the environment. Opposed to this is the advantage of an ecologically-valid naturalistic environment in terms of driver motivation (Smiley & Brookhuis, 1987). A crash in a laboratory test or simulator has no serious consequences. Out on the road, however, not many collisions can be afforded. Although the driver's motivation is higher in a field test, there still are differences compared with normal driving. Demand characteristics and the presence of an experimenter who can handle redundant controls in case of an emergency, cannot be excluded as having an effect on the driver's behaviour. Also, it should be clear that there is more to workload than task parameters alone: for example the processing of task-irrelevant information and emotional information have an effect on mental workload (see Meijman & O'Hanlon, 1984). Moreover, when performing traffic research, not only measurement techniques have to be carefully chosen. Equally important is selection based upon representativeness of subjects, variables and setting. For an overview of these parameters see Kantowitz (1992b).

Driver mental workload can be affected in many ways. An affected driver state caused by monotony can become overt in driving-performance parameters. In the case of increased task complexity these primary-task performance parameters will also be affected. Other measures will be differentially affected by these two factors that increase workload. Some of the physiological measures are more sensitive to increased task complexity than to reduced driver state. Very important in mental workload research are the two areas in which the driver is compensating for altered demands by increasing effort. Performance parameters in general will not indicate the additional costs to the driver, while other measures, such as self-report and physiological measures, may. For this reason the major conclusion is that in experimental research the use of a single measure of workload is not sufficient for the assessment of driver mental workload. The different studies discussed support this view.

The psychological concepts that are used in mental workload research have been differentially defined in different studies. Resources and capacity are used as interchangeable terms as is the case with complexity and difficulty, and 'load' is used to indicate cause and effect. Nevertheless, even if one sticks to a definition of a concept, and workload is defined as the reaction to task demands, then individual

task goals that can and are set in the field will diffuse these demands between individuals. While laboratory tasks are often well defined, new and amazingly simple (e.g., press a button when you hear a tone), ecologically valid tasks, such as car driving, are diffuse or are composed of several subtasks, are complex and well-trained. It seems safe to state that strategies and automation in performance of subtasks play a large role in behaviour that is frequently displayed (read: in driving) opposed to infrequently displayed behaviour (read: laboratory tasks). The self-pacedness of driving makes compensatory behaviour possible, thus leaving a part of the regulation of task demands in the drivers' hands. All these aspects are also present in the measurement of driver mental workload, and probably in any applied setting. Nevertheless, many of the measures that were developed and first tested in the laboratory turned out to be very sensitive in the field. The 0.10 Hz component of heart rate variability is a good example of a measure that can be used in driving a car, and is very sensitive to mental effort. Results show that the measure is sensitive to task-related effort, thus supporting the idea that it reflects the defense response. Perhaps the restricted space for making physical movements and the overall activating effect of car driving places the subjects in an ideal 'state' to measure differences in mental effort on this parameter. While heart rate can be registered in a car and is very useful, some of the other measures are difficult or impossible to register. Pupillometry is not useful in traffic research in the field due to changes in ambient lightning and most secondary tasks distort primary-task performance.

In my view, basic and applied research can benefit from each others' knowledge and experience. The laboratory is a environment in which workload measures can be developed and tested with tasks that can be controlled to a large extent. In the field, the measure's sensitivity can then be further assessed using well-practised tasks in which goal setting also plays a very important role. Results should be fed back to the laboratory for evaluation and possible improvement of the measures. Both basic and applied research contribute to the understanding of the processes involved in mental workload and both types of research need each other. Without basic research, very few of the advanced measures would have been developed, while applied research maintains that 'the proof of the pudding is in the eating', and should be carried out in situations that approach the complexity of everyday life.

