The measurement of drivers' mental workload

de Waard, Dick

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
1996

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Measures

O’Donnell & Eggemeier (1986) specify three workload-measurement groups: subjective (i.e., self-report) measures, performance measures and physiological measures. All categories will be considered separately below. Performance measures are split into three categories: primary-task performance measures, secondary-task performance measures and reference tasks. An overview of most measures will be given, although some of the measures will receive more attention than others. The reason for this is that these measures will be evaluated in chapter 5 on their use in traffic research. Evaluation will focus on use of the measures as indicators of mental load in case of an affected driver state opposed to sensitivity to increases in task complexity.

4.1 Self-report measures

Self-report measures have often been indicated as subjective measures. The reason for preferring the word ‘self-report’ to ‘subjective’ is that measures from other measurement groups, in particular physiological measures, are also subjective (see also Muckler & Seven, 1992). Self-report measures have always been very appealing to many researchers. No one is able to provide a more accurate judgement with respect to experienced mental load than the person concerned. Sheridan (cited in Wickens, 1984) considers self-report measures to be the best measures since they come nearest to tapping the essence of mental workload. Critics, on the other hand, say that the source of the resource demands is hard to introspectively diagnose within a dimensional framework. Physical and mental workload are, according to the critics, hard to separate (see e.g., O’Donnell & Eggemeier, 1986).

Muckler & Seven (1992) state that the strength of self-report measures is their subjectivity. "The operator’s awareness of increasing effort being used, even before any performance degradation occurs, should give subjective [self-report] measures a special role to play". Different dimensions of workload, such as performance and effort, are integrated in self-report measures while at the same time individual differences, operator state and attitude are taken into account. According to Muckler & Seven (1992) these differences are obscured in objective measures until breakdown makes them obvious in performance measures. This last statement may be true for primary-task performance, it does not hold for some of the physiological measures and/or dual-task performance (see 4.2 and 4.3).

Most self-report measures are sensitive in all but the A2 region. In the A1 and A3-region ratings of effort could indicate the increase in workload. In the C-region severe overload occurs which
could become apparent from low performance combined with high activation-ratings, or ‘quitting’ behaviour.

**RSME, Rating Scale Mental Effort**

In the Netherlands, a unidimensional scale, RSME (Rating Scale Mental Effort), was developed by Zijlstra (Zijlstra & Van Doorn, 1985, Zijlstra & Meijman, 1989, Zijlstra, 1993). Ratings of invested effort are indicated by a cross on a continuous line. The line runs from 0 to 150 mm, and every 10 mm is indicated. Along the line, at several anchor points, statements related to invested effort are given, e.g., ‘almost no effort’ or ‘extreme effort’ (see appendix A). The scale is scored by measurement of the distance from the origin to the mark in mm. On the RSME the amount of invested effort into the task has to be indicated, and not the more abstract aspects of mental workload (e.g., mental demand, as is in the TLX, see below). These properties make the RSME a good candidate for self-report workload measurement.

**Activation scale**

On the unidimensional activation\(^2\) scale (Bartenwerfer’s scale, Bartenwerfer, 1969) subjects are required to mark a line. The looks of the scale are comparable to the RSME, the activation scale also consists of a single axis with reference points on it. However, at the reference points statements of a different nature are given, like ‘I’m reading a newspaper’ and ‘I am trying to cross a busy street’ (see appendix B). Subjects are asked to mark the line with a cross at the position that equals their mental activation during task performance. The scale has a range from 0 to 270 and is scored by measuring the distance from the origin to the mark in millimetres.

**Other self-report measures**

Three frequently used rating scales are the NASA Task Load Index (TLX, Hart & Staveland, 1988), the Subjective Workload Assessment Technique (SWAT, Reid et al., 1981) and the Modified Cooper-Harper scale (MCH, Wierwille & Casali, 1983). Both the TLX and the SWAT are multidimensional scales. This means that ratings on several subscales (e.g., scales regarding experienced time-pressure, physical load) have to be completed. In the end these ratings can be summarized to obtain an overall workload assessment. In order to obtain an overall workload rating with the TLX, first the six scales should be compared to each other for each task and the operator has to rate which of the two dimensions contributed most to his or her feeling

\(^2\) The word activation as used here has a broader meaning than the concept of activation as used by Pribram & McGuiness (1975). Here the word activation covers experienced mental activation as well as feelings of arousal.
of workload. This necessitates a total of 15 comparisons before the overall workload rating can be calculated. The MCH is a unidimensional scale in which a series of questions directly lead to a single rating. For an overview of these three rating techniques and a comparison of their sensitivity in non-aviation field settings, see Hill et al. (1992). They concluded that the TLX and a fourth, less common and unidimensional scale (‘Overall Workload scale’) were the best measures with respect to sensitivity to workload. Veltman and Gaillard (in press) compared the NASA TLX multidimensional scale with the RSME in an experiment using a flight-simulator. They found that the RSME was more sensitive than the TLX. The authors argue that this result may be related to confusion caused by the TLX-subscales.

While the ‘traditional’ TLX requires a two-pass process with paired comparisons, Byers et al. (1989) have proposed a Raw Task Load Index (RTLX) which does not require task paired comparison weights. The RTLX is a simple average of the six TLX scales. Byers and his colleagues found that TLX and RTLX had comparable means and standard deviations, and correlated above $r = 0.95$, and they recommend the RTLX as a simple alternative to the TLX. These findings are supported in a report by Fairclough (1991).

**Unidimensionality versus multidimensionality**

Which rating scale to use depends on what information is needed. Diagnosticity is probably larger for multidimensional scales (Nygren, 1991, Hill et al., 1992). If, however, a global rating of workload is required, then the subject’s univariate workload rating is expected to provide a measure that is more sensitive to manipulations of task demands than is a scalar estimate derived from judgements along several individual workload-related factors (Hendy et al., 1993). Muckler & Seven (1992) also stress the simplicity self-report scales should have. If possible the measures should have immediacy and be comprehensible to reduce the need for interpretation and to aid in the precision of measure definition. This is mainly true for unidimensional scales.

Unidimensional scales can be given multidimensional properties if they are applied separately per task-dimension. Zijlstra and Meijman (1989) have used the RSME in this way; they asked people to rate different dimensions of task performance separately. In this study a RSME rating was obtained by rating the effort required to perform different sub-tasks, such as navigation, machine-use and communication. The advantage of this method is that a more differentiated picture emerges. It can be argued however, that multiple use of a unidimensional scale in this way is not fundamentally different from multidimensional scales.

Self-report scales have several advantages, the major advantage perhaps being their high face validity. In addition, the ease of application and low costs can be mentioned. Low primary-task intrusion
is secured as long as the scale is administered after completion of the task. Delays of up to 30 minutes in workload reporting do not lead to significant differences, with the possible exception of delayed ratings after complex multiple-task performance (Eggemeier & Wilson, 1991). Other limitations of self-report measures include (see O’Donnell & Eggemeier, 1986) a possible confusion of mental and physical load in rating, the operator’s inability to distinguish external demands from actual effort or workload experienced. O’Donnell & Eggemeier (1986) also consider a possible dissociation between self-report measures and performance to be an aspect that restricts use. Also mentioned are limitations in the operator’s ability to introspect and rate expenditure correctly, which, e.g., become obtrusive in conflicting findings in that either peak workload or average workload level determine the final rating (e.g., Vidulich & Tsang, 1986).

4.2 Performance measures

Primary-task measures

In laboratory tasks, motor or tracking performance, the number of errors made, speed of performance or reaction time measures are frequently used as primary-task performance measures. Outside the laboratory, primary-task performance is, by its nature, very task-specific. There is not one prevalent primary-task measure, although all primary-task measures are speed or accuracy measures.

According to O’Donnell & Eggemeier (1986) primary-task performance is a measure of the overall effectiveness of man-machine interaction. As discussed under sensitivity (chapter 3) there are some limitations to this statement. Primary-task performance diminishes outside the A region, while a constant performance in the A region does not necessarily reflect low operator workload. No performance differences between two operators can be determined, even though one can be ‘at the limit of his capability’, while the other is capable of performing an additional task, without any change in primary-task performance level. Therefore it is necessary to combine primary-task performance and other workload measures in order to draw valid conclusions about man-machine interaction and, in particular, about the operator’s strategy or energetic state.

Secondary-task measures

When another task is added to the primary task, secondary-task measures can be taken. Two paradigms can be applied to dual-task performance (see O’Donnell & Eggemeier, 1986). Within the ‘Loading Task Paradigm’ secondary-task performance is maintained, even if decrements in primary-task performance occur. The addition of the second task results in a total workload shift from region A towards region B, so that primary-task performance measures can be used as
indicators of workload. Within the second paradigm, the ‘Subsidiary Task Paradigm’, the instruction to maintain primary-task performance is given. Consequently secondary-task performance varies with difficulty and indicates ‘spare capacity’, provided that the secondary task is sufficiently demanding. Spare capacity (Brown & Poulton, 1961) is a concept that is used frequently in dual task performance, and assumes a total undifferentiated capacity that is available to perform all tasks. In the case of unaffected single-task performance, the unused capacity is called spare capacity, and is in principle available for secondary-task performance.

According to the multiple-resource theory (Wickens, 1984) the largest sensitivity in secondary-task measures is achieved if the overlap in resources that are used is high. In other words, in order to perform the secondary task, spare capacity of the same resource should be required. Time sharing is expected to be less efficient if the same resources are used. This large overlap in resources used is at the same time a threat to undisturbed primary-task performance because primary-task intrusion is largest if two tasks that use the same resources have to be time-shared. Other problems that are related to secondary task methodology (Eggeemeier & Wilson, 1991) are non-specific intrusion (e.g., peripheral interference), the omission of secondary-task performance in the case that primary-task demands are very high, and the operators’ resource allocation policy (the priority given to each task). This resource allocation policy is in particular important if the primary task has a high ecological validity. Also, the choice for a secondary task is more difficult in tasks approaching everyday performance. Car driving, for instance, is to a large extent automated and mainly a visual task. The value of a secondary auditory digit-addition task is therefore not completely distinct. It is possible that performance on the latter task reflects central resource use. However, the extent to which performance of the primary task makes use of central resources is not clear in advance. The use of secondary tasks in applied environments is more complex than in laboratory experiments, and for this reason caution is required.

Most frequently used as secondary tasks are choice reaction-time tasks, time estimation or time-interval production, memory-search tasks and mental arithmetic (see O’Donnell & Eggeemeier, 1986, Eggeemeier & Wilson, 1991 and Wickens, 1992, for overviews). Eggeemeier & Wilson (1991) have compared several multiple-task studies and conclude that results regarding sensitivity of the different measures are mixed. Primary-task intrusion also differs between studies. They argue that both effects are related to a large diversity in workload.
levels, tasks and test environments. Relatively low primary-task intrusion is to be expected with the irrelevant-probe technique. As general disadvantages of secondary-task techniques, Eggemeier & Wilson (1991) mention: the requirement of additional instrumentation, possible compromises to system safety (primary-task intrusion) and a lack of operator acceptance. Some of these problems are overcome if embedded secondary-task measures are used. An embedded secondary task is ‘an operator function performed during normal system operations, but distinct from the primary operator function that is under assessment’ (Eggemeier & Wilson, 1991). The priority assigned to these tasks is lower than that assigned to the primary function, and thus primary-task intrusion is expected to be limited. As embedded tasks are part of the operator’s role in the system environment, operator acceptance is high. Also, the embedded task itself is not artificial. An example of an embedded task is the number of radio communications, or the length of communications, that occur during a flight. A relatively new alternative could be secondary-task performance in terms of speech measures. As a secondary counting task (counting from 90 to 100), speaking fundamental frequency (pitch), speaking rate and vocal intensity (loudness) have been found to be sensitive indicators of workload (Brenner et al., 1994). A major advantage of speech measures is that the collection of the indices itself is unobtrusive and no equipment has to be attached to the subject. However, the secondary-task technique in the above-mentioned format is by no means unobtrusive. If normal speech could be used instead of a secondary (counting) task, then an embedded task would emerge and that would mean a large step forward. As the differences in the speech measures found in the laboratory were small in absolute value there is unfortunately little reason to expect that ordinary conversation speech measures can be used as workload indicators in the near future.

Reference tasks

Reference tasks are listed here for the sake of completeness. Reference tasks are standardized tasks that are performed before and after the task under evaluation and they mainly serve as a checking instrument for trend effects. Changes in performance on reference tasks can indicate effects of mental load of the primary task. If subjective and physiological measures are added to the reference tasks the costs for maintaining performance on the primary task could also be inferred, in particular if the operator’s state is affected. The use of standard

\footnote{In the irrelevant probe technique subjects do not need to respond to auditory-presented stimuli. However, certain brain potentials are evoked by these stimuli and their amplitude and latency can be indicative of mental load. Due to this main measure the technique is discussed further under physiological measures (section 4.3).}
reference task batteries is very common in organizational and occupational psychology (see, e.g., Van Ouwerkerk et al., 1994b).

4.3 Physiological measures

The last category of workload measures are those derived from the operator’s physiology. Different physiological measures have been found to be differentially sensitive to either global arousal or activation level (e.g., pupil diameter), or to be sensitive to specific stages in information processing (e.g., the evoked cortical brain potential). The advantage of physiological responses is that they do not require an overt response by the operator, and most cognitive tasks do not require overt behaviour. Moreover, most of the measures can be collected continuously, while measurement is nowadays relatively unobtrusive due to miniaturisation. Kramer (1991) mentions as disadvantages of physiological measures the required specialized equipment and technical expertise, and the critical signal-to-noise ratios. He also states that the operator’s physiology, a reflection of bodily functions, is further removed from operator-system performance than, e.g., primary-task performance.

Measures from two anatomical distinct structures are used as physiological indicators, Central Nervous System (CNS) measures and Peripheral Nervous System measures. The CNS includes the brain, brain stem and spinal cord cells. The Peripheral Nervous System can be divided into the Somatic Nervous System and the Autonomic Nervous System (ANS). The Somatic Nervous System is concerned with the activation of voluntary muscles, the ANS controls internal organs and is autonomous in the sense that ANS innervated muscles are not under voluntary control. The ANS is further subdivided into the Parasympathetic Nervous System (PNS) and the Sympathetic Nervous System (SNS). While the PNS function is to maintain bodily functions, the SNS function is directed towards emergency reactions (see, e.g., Matsumoto et al., 1990, Kramer, 1991). Most organs are dually innervated, i.e., both by the sympathetic and the parasympathetic nervous systems. While traditionally these branches are seen as subject to reciprocal central control -as a continuum from parasympathetic to sympathetic dominance- recently, a two-dimensional autonomic space was proposed with a parasympathetic and a sympathetic axis (Berntson et al., 1994). SNS and PNS can be coactive, reciprocally active, or independently active. Some evidence for autonomic space was provided in the same paper (Berntson et al., 1994).

Examples of ANS-measures are the pupil diameter, heart rate and respiratory, electrodermal and hormone level measures. CNS-measures include electrical, magnetic and metabolic activity of the brain and electrooculographic activity. A third category of measures are
peripheral responses that include spontaneous muscle activity and eye movements (see O’Donnell & Eggemeier, 1986).

Overviews of physiological measures of workload are given by O’Donnell & Eggemeier (1986) and Kramer (1991). Emphasis here is on measures that can be used outside the laboratory, in particular in traffic research. Where possible an update on the above-mentioned overviews will be provided.

**Cardiac Functions**

The heart is innervated both by the PNS and the SNS and each heart contraction forces the blood through the circulatory system. The contraction is produced by electrical impulses that can be measured in the form of the ECG (ElectroCardioGram). From the ECG signal (a) time domain measures, (b) frequency measures and (c) amplitude measures can be derived.

In the time domain the R-waves (see, e.g., Kramer, 1991, L.J.M.Mulder, 1992) of the ECG are detected, and the time between these peaks, the Inter-Beat-Interval (IBI), is calculated. **Heart Rate (HR)** is directly related to Heart Period (HP) or IBI⁴, however, this relation is non-linear and IBI is more normally distributed in samples compared with HR (Jennings et al., 1974). Therefore, IBI scores should be used for detection and testing of differences between mean HR scores, the IBI scale is less influenced by trends than the HR scale (Heslegrave et al., 1979). Average heart rate during task performance compared to rest-baseline measurements is a fairly accurate measure of metabolic activity (Poros & Byrne, 1992). Roscoe (1992) claims that the main determinant in heart rate response in experienced pilots, in the absence of physical effort, is workload. However, pilot workload levels are probably higher than workload levels in laboratory experiments or in automobile driving (cf., selection criteria for pilots vs. driving-licensing criteria, and see also Wilson, 1992). Not only physical effort affects heart rate level (e.g., Lee & Park, 1990), emotional factors, such as high responsibility or the fear of failing for a test, also influence mean heart rate (Jorna, 1993). Other factors affecting cardiac activity are speech and high G-forces (Wilson, 1992). The effect of sedative drugs and time-on-task resulting in fatigue is a decrease in average HR (e.g., Mascord & Heath, 1992), while low amounts of alcohol are reported to increase HR (e.g., Mascord et al., 1995).

A continuous feedback between the CNS and peripheral autonomic receptors causes irregularities in heart rate. Heart rate variability is a marker of performance of this feedback system and in

---

⁴ Heart Period is expressed in cardiac time (‘beat-by-beat’), while Heart rate is expressed as count of beats per ‘real’ time (see Papillo & Shapiro, 1990). HR = 60,000 / IBI, HR in BPM -Beats Per Minute-, IBI in milliseconds
healthy humans this is reflected in large deviations from the mean rate (e.g., Porges, 1992). *Heart Rate Variability (HRV)* in the time domain is also used as measure of mental load (Kalsbeek & Ettema, 1963). If HRV is referred to as variability coefficient or modulation index, the measure is standardized by dividing the standard deviation of IBIs by the average IBI. HRV provides additional information to average HR about the feedback between the cardiovascular systems and CNS structures (see Porges & Byrne, 1992). In general HRV decrease is more sensitive to increases in workload than HR increase, although there have been several reports of both HR and HRV insensitivity (e.g., Wierwille et al., 1985). One of the causes for finding no effect of mental load on HRV lies in the globalness of the measure and its sensitivity to physical load. Lee & Park (1990) showed that an increase in physical load decreased HRV and increased HR, while an increase in mental load was accompanied by a reduced HRV and no effect on HR. Fatigue is reported to increase HRV (Mascord & Heath, 1992) while low amounts of alcohol decrease HRV (Gonzalez Gonzalez et al., 1992). Mascord et al. (1995), however, report an increase in HRV as a result of low amounts of alcohol and attribute this to alcohol-induced fluctuations in the autonomic control of heart rate.

Compared to time-domain analysis, frequency analysis of IBI has as a major advantage that HRV is decomposed into components that are associated with biological control mechanisms (Kramer, 1991, Porges & Byrne, 1992). Three frequency bands have been identified (see L.J.M.Mulder, 1988, 1992): a *low frequency* band (0.02 - 0.06 Hz) believed to be related to the regulation of the body temperature, a *mid frequency* band (0.07 - 0.14 Hz) related to the short-term blood-pressure regulation and a *high frequency* band (0.15 - 0.50 Hz) believed to be influenced by respiratory-related fluctuations (vagal, PNS influenced, see Kramer, 1991). A decrease in power in the mid frequency band (also called the ‘0.10 Hz component’ after the main frequency component), and in the high frequency band have been shown to be related to mental effort and task demands (G.Mulder, 1980, Mulder & Mulder, 1981, Aasman et al., 1987, Vicente et al., 1987, L.J.M.Mulder, 1988, Itoh et al., 1990, Jorna, 1993, Veltman & Gaillard, 1993, Backs & Seljos, 1994). Jorna (1992) and Paas et al. (1994), however, conclude that spectral measures are primarily sensitive to task-rest differences, and not to moderate increases in difficulty within a task. According to Jorna (1992) only large differences, such as the transition from single to dual task or automatic vs. controlled processing, are able to induce observable differences on spectral measures. It might also be that, instead of being sensitive to major differences in task load, the 0.10 Hz component is most sensitive in relatively low workload areas. In the higher workload regions, the areas where performance is affected to a great extent and overload emerges, the measure’s sensitivity is non-linear to workload increases (cf. Aasman et al., 1987).
Finally, amplitude information from the ECG signal can be utilized to obtain information about workload. The amplitude of the T-wave (TWA) is said to mainly reflect SNS activity (Furedy, 1987) and decreases with increases in effort. Some support for sensitivity in terms of a TWA decrease with increases in SNS activity, as well as for PNS-activity influence on respiratory sinus arrhythmia, is provided by Müller et al. (1992). In table 2 alternative naming of heart rate measures and HRV-frequency bands are listed.

<table>
<thead>
<tr>
<th>Variable/Frequency band</th>
<th>Abbreviation</th>
<th>Alternative name, ¹ = inverse (related)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heart Rate</td>
<td>HR</td>
<td>Inter-beat-interval (IBI)¹, Heart Period (HP)¹</td>
</tr>
<tr>
<td>Heart Rate Variability</td>
<td>HRV</td>
<td>Sinus Arrhythmia, Variation coefficient (Modulation index)</td>
</tr>
<tr>
<td>T-wave</td>
<td>TWA</td>
<td>T-wave Amplitude, Temperature band, Slow-wave component</td>
</tr>
<tr>
<td>Low frequency band</td>
<td>.10 Hz</td>
<td>0.10 Hz band, 0.10 Hz component, Blood pressure band, T-H-M-Wave (Traube-Hering-Mayer)</td>
</tr>
<tr>
<td>Mid frequency band</td>
<td>RSA</td>
<td>Respiratory Sinus Arrhythmia, 'V'-component (vagal), Respiration band</td>
</tr>
<tr>
<td>High frequency band</td>
<td>RSA</td>
<td>Respiratory Sinus Arrhythmia, 'V'-component (vagal), Respiration band</td>
</tr>
</tbody>
</table>

Measurement of heart rate is not very complex, the ECG signal needs little amplifying (about 10 to 20 times less as ongoing EEG) and if measurement is limited to R-wave detection and registration then electrode placement is not very critical. Heart rate may provide an index of overall workload, spectral analysis of heart rate variability is more useful as index of cognitive, mental workload (Wilson & Eggemeier, 1991). A restriction in the use of heart rate measures is that, due to the idiosyncratic nature of the measure, operators are usually required to serve as their own control in workload assessment. Another major restriction to the use of ECG measures is the effect speech has on blood pressure, and therefore on the 0.10 Hz component of heart rate variability (L.J.M.Mulder, 1988, Sirevaag, 1993). If verbalization is a predominant aspect of operator performance the 0.10 Hz component may be less suitable for mental load assessments. However, speech is not necessarily a disturbing factor, Porges & Byrne (1992) recommend no corrective action in cases in which the verbalization duration is short (less than 10 s) or in the case that speech is relatively infrequent (one to five times per minute). Another important factor influencing HRV is physical load. The 0.10 Hz frequency component, however, has been shown to be relatively insensitive to light physical load (e.g., Hyndman & Gregory, 1975, Fairclough, 1993). Also, if physical load is not extreme and it is kept constant across conditions, the 0.10 Hz
component of HRV may well be used to indicate mental effort. Finally, age may affect the use of HR measures, restriction of subjects to specific age groups may be required if HRV is the primary workload measure. HRV may decrease with increasing age due to, amongst others, a decrease in blood vessel flexibility (G. Mulder, 1980). With elderly subjects, the measure may turn out to be less sensitive than expected.

In the 1980’s relatively long data time windows of at least 100 seconds had to be used for spectral analysis. In this decade, advanced techniques have become available, such as profile analysis (L.J.M.Mulder et al., 1990) that can use smaller time windows of, e.g., 30 s, and the COMMOD technique (COMplex deMODulation, see Jorna, 1993), which digitally filters the HR signal in a selected frequency band. With the aid of these techniques, changes in HR and HRV during the course of task performance can be monitored.

**Background Electroencephalogram (EEG)**

An electroencephalogram is a recording of electrical activity made from the scalp. Frequency analyses performed on the EEG signal are typically classified into the following ranges or bands (see, e.g., Cooper et al., 1980):

- up to 4 Hz: Delta waves,
- 4 to less than 8 Hz: Theta waves,
- 8 to 13 Hz: Alpha waves and
- more than 13 Hz: Beta waves

Frequency analyses are also referred to as epoch analyses, or background EEG analyses and reflect tonic CNS activity. Delta rhythms are present during deep sleep while beta waves predominate during active wakefulness. In general alpha and theta waves are associated with decreased alertness, though individual differences may be large. There is, for instance, a minority of people who do not generate alpha waves at all.

Epoch analysis on EEG in mental workload research is rare and less common than EEG spatial pattern analysis (see section on ERPs under ‘Other measures’). In the workload studies in which EEG frequency analyses were calculated, in general alpha and theta sensitivity is reported (Kramer, 1991). Sirevaag et al. (1988) report a decrease in alpha activity and an increase in theta during dual-task performance opposed to single-task performance. The use of EEG frequency analysis is, however, far more customary in operator state assessment, e.g. the assessment of arousal level during vigilance situations (Wilson & Eggemeier, 1991). Clearly, more research regarding the relation between background EEG and mental workload - and in particular the relation with increased task complexity - is needed.
to be able to judge the measure on its usefulness as indicator of mental workload.

**Eye fixations**

Some measures are hard to classify as either performance or physiological measures. An example of such a measure are measures of eye fixations. Eye fixations are related to primary-task performance (most tasks are of a highly visual nature). Eye fixations could be considered secondary-task performance measures in the case of embedded tasks (e.g., when the secondary task is to monitor an additional device), but traditionally fixations are listed under physiological parameters, probably due to one of the measurement techniques, the ElectroOculoGram.

Visual-search strategy, or the selective attention to relevant visual stimuli, has been shown to be indicative of information needs (Hughes & Cole, 1988). The eye-scanning patterns of pilots in terms of frequency of fixation were found to be related to instrument importance. The length of fixations, however, was related to difficulty in obtaining/interpreting information from instruments (see Wilson & Eggemeier, 1991). O’Donnell & Eggemeier (1986) report that an increase in workload is accompanied by increased fixation time. Backs & Walrath (1992) also determined fixation time (‘dwell time’) in a visual high-demand situation. They found that fixation time differed depending upon task characteristics. An increased fixation time was found in self-terminating search vs. exhaustive search, and increased fixation time was also found for stimuli that were monochrome opposed to colour coded. Backs & Walrath (1992) explained this dependency in terms of differences in participant strategy.

When a precise fixation is required, or in a tracking task, the size of the functional field of view may indicate processing demands. The functional field of view (Sanders, 1970) is an area around the central fixation point from which information is actively processed during performance of a visual task. May et al. (1990) report a significant decrease in the range of saccadic extent as a result of mental workload in a laboratory task. With an increase in load the saccadic range decreased.

The main problem with eye point-of-regard analysis is that eye fixations always ‘fill up’ the total time. This is in particular a problem in low to moderate workload situations, in which not all fixations are relevant and required for task performance. Moreover, the sensitivity of measures of eye-fixation will be restricted to visual workload, and the measure can be considered diagnostic in that respect. Another problem related to ‘filling up of fixation time’, is the difference between looking and perceiving. A fixation does not necessarily imply perception.

Eye fixations can be measured using video camera registration, by registration of cornea reflection superimposed on a video image of the visual field, or by the registration of the ElectroOculoGram (EOG).
The EOG technique has as a disadvantage that an accurate foveal point-of-regard is hard to assess. The video techniques both suffer from labour-intense and time-consuming data analysis. The cornea reflection technique is accurate in point-of-regard evaluation, as long as the equipment is calibrated regularly, i.e. every 15 minutes or so. An advantage of modern equipment is that it is no longer head-mounted, which minimizes primary-task intrusion. Nevertheless, the measurement of eye movements of subjects wearing glasses is very difficult.

**Other physiological measures**

**Pupil diameter**

Pupil diameter decreases as a result of activity of PNS-innervated muscles, while SNS-innervated muscle groups cause a pupil dilation. Kahneman put pupil diameter forward in his book *Attention & Effort* (Kahneman, 1973) as an important measure of mental workload. He concluded that increased task processing demands and increased resource investment were reflected in increases in pupil diameter. Beatty (1982) reports the same relationship between mental workload and pupil diameter: pupil diameter increases with increases in perceptual, cognitive and response-related processing demands. As most arousal-related measures, the pupil diameter as measure is not diagnostic and has been used as an indicator of global workload. Backs & Walrath (1992) give the following description of stimulus-related pupillary response measurements. In a single-trial the pupillary response shows two components. After baseline a large constriction-peak follows about 950 ms after stimulus onset. This is followed by a gradual dilation peaking dependent upon search time. Peak-to-peak differences between the two components are used after baseline subtraction. In their study (Backs & Walrath, 1992) subjects had to search visual displays. The effects they found in pupillary response were related to information-processing demands. Recently, the pupil diameter has received renewed interest. Hoeks (1995) and Hyöna et al. (1995) have published studies in which the pupillary response was related to mental processing load, while Wilhelm & Wilhelm (1995) linked low frequency ‘pupillary oscillations’ to fatigue.

Even though effects of mental load on pupillary response were found, the largest changes in pupil diameter occur as a result of other factors, e.g., a change in ambient illumination and the near reflex. These factors make the measure best suitable for laboratory situations (Kramer, 1991).

**Endogenous eye blinks**

Endogenous eye blinks, i.e. eye blinks in the absence of an identifiable eliciting stimulus, can be measured by corneal-reflection techniques, video scanning or electrooculogram (EOG). The sensitivity to workload of three components of eye blink has been studied, (a) *eye blink rate*, (b) *blink duration* and (c) *eye blink latency*, the latter measure in relation to stimulus occurrence. Kramer (1991) states in his
review that results related to blink rate are mixed, while latency increases and closure duration decreases with increases in task demands. Stern et al. (1994) conclude that increased blink frequency is a meaningful reflector of fatigue. When measuring eye blink duration the EOG measurement technique is more reliable than video. Due to video resolution short-lasting blinks (20-30 ms) could be missed (Wilson & Eggemeier, 1991). Eye functions seem most useful in assessment of visual demands, and not in auditory or cognitive demand situations (Kramer, 1991, Sirevaag et al., 1993). Just as pupil diameter, selectivity of eye blinks to workload is low. Other factors than workload, e.g., the quality of the air quality, affect blink measures.

**Blood pressure**

Closely related to a decrease in HRV is the decrease in blood-pressure variability (BPV). If a decrease in HRV is caused by a decrease in baroreflex sensitivity then this will be reflected in reduced BPV (see G.Mulder, 1980, L.J.M.Mulder, 1988). Continuous blood-pressure measurements are required to demonstrate BPV. These measurements are accomplished by enclosing a finger in a small cuff. The cuff is either filled with water (Steptoe & Sawada, 1989) or with air (FIN.A.PRES, Settels & Wesseling, 1985). The pressure in the cuff is adjusted to intra-arterial blood pressure and can be monitored. The technique is, however, best fit for the laboratory and it has been applied there successfully in mental load tasks (see, e.g., L.J.M.Mulder, 1988).

**Respiration**

Respiration is indispensable to supply the blood with oxygen and to expel carbon dioxide. Measures of respiration could provide an index of energy expenditure. Recently, evidence has been found supporting the hypothesis that cognitive effort coincides with a small but significant increase in energy expenditure (Backs & Ryan, 1992, Backs & Seljos, 1994). The most frequently used measure of respiration is respiration rate (Wilson & Eggemeier, 1991). Respiration rate increases under stressful attention conditions (e.g., Porges & Byrne, 1992) and as a result of increased memory load or increased temporal demands (Backs & Seljos, 1994). Wientjes (1992, 1993) states that respiration rate without information about tidal volume is meaningless and has led to inconclusive results. The multiplication of respiration rate (i.e., timing) with tidal volume (i.e., intensity) gives the minute ventilation, the quantity of air breathed per minute. Wientjes (1993) found an increase in minute ventilation (and an increase in respiration rate and a decrease in tidal volume) as a result of mental effort or mild stress.

The main problems with respiration measures are related to the measurement technique. Accurate flow meters can be used that can analyze expired gasses, but these devices add dead space and resistance, and are very intrusive. Indirect measurement techniques such as strain gauges, impedance pneumography and equipment that measures
changes in air flow temperature, may be less intrusive, but these
techniques are also less accurate (for a discussion of techniques, see
Porges & Byrne, 1992). Wientjes (1993) reports a method that is both
non-invasive and provides time and volume information. The method
assesses separate rib cage and abdomen motions. However, at certain
intervals calibration sessions with the previously mentioned flow meters
are required or, alternatively, subjects have to breathe a fixed known
volume. This clearly makes the technique, compared to for instance
ECG measurement, more complicated. Moreover, the measure is, just as
many other physiological measures, not uniquely sensitive to mental
effort and is affected by, for instance, speech and physical effort. It is
also closely linked to emotions and personality characteristics. Wientjes
(1992) as well as Backs & Seljes (1994), however, consider the use of
respiration measures to be undervalued in psychophysiological research.
In applied settings, respiration measures, in particular respiration rate,
have been used several times as measures of mental load. Use of the
measures has been confined to aviation, mainly to (simulated) high-
speed jet-flight (see, e.g., Roscoe, 1992, Wilson, 1992). In these field
studies it was also found that, in general, a decrease in respiration rate
coincided with increases in cognitive activity.

Electrodermal Activity, EDA

Electrodermal activity refers to the electrical changes in the
skin. These changes are the result of ANS activity. Two techniques are
in use, exosomatic and endosomatic measurement. With exosomatic
measurement a small current from an external source is led through the
skin and is measured, while the less frequently applied endosomatic
measurement makes no use of an external source. EDA is expressed in
terms of skin conduction or resistance, which are (nonlinearly) inversely
related. Electrodermal activity can be further distinguished in tonic and
phasic activity (Heino et al., 1990), while Kramer (1991) adds
spontaneous or non-specific EDA to these two. Tonic EDA, the
Electrodermal Level (EDL) or Skin Conduction Level (SCL), is the
average level of EDA or baseline activity. Phasic EDA includes the
Electrodermal Response EDR, which is most similar to the formerly
common measure GSR (Galvanic Skin Resistance). EDR is the result of
an external stimulus. Response is fairly slow, a latency of 1.3 to 2.5 s
to the occurrence of stimulation is to be expected (Kramer, 1991). EDR
is expressed either as Skin Resistance Response (SRR) or as Skin
Conduction Response (SCR).

Spontaneous EDA, EDA in response to unknown stimuli, has
predominantly been used as an indicator of arousal or emotion, and not
as a measure of workload. Kramer (1991) in his review, refers to
several studies that show sensitivity of SCR to information processing.
He concludes that spontaneous EDA appears to be sensitive to general
levels of arousal while SCRs seem to index the allocation of an
undifferentiated form of processing resources. The main problem with
electrodermal activity measures are a global sensitivity, or as Heino et
al. (1990) state “all behaviour (emotional as well as physical) that affects the sympathetic nervous system can cause a change in EDA”. EDA is usually measured on the palm of the hand or on the sole of the foot where SNS-controlled eccrine sweat glands are most numerous (Dawson et al., 1990, Kramer, 1991). Activity of these glands is sensitive to respiration, temperature, humidity, age, sex, time of day, season, arousal and emotions. The measure is therefore not very selective.

**Hormone levels**

Certain hormones are released under SNS-stimulation in stress situations, which includes high workload situations (Wilson & Eggemeier, 1991). Of particular interest are the catecholamine hormones Adrenaline (A) and Noradrenaline (NA). The adrenal cortical steroid Cortisol is also frequently used as a stress indicator. Hormone levels reflect integrated effects of stress over time and can be measured from urine samples, blood or saliva. An increase in time to return to baseline values or an elevated hormone level may provide an indication of workload (Meijman & O’Hanlon, 1984). Increased NA and A levels occur in cases of effortful coping (e.g. Meijman, 1989, Van der Beek et al., 1995). If, apart from increased A and NA levels, cortisol levels are also increased, and these levels remain elevated for longer periods of time, then the operator is in a state of ‘effortful distress’ (Frankenhaeuser, 1989, Van Ouwerkerk et al., 1994b).

With respect to sensitivity, there is evidence that separation of mental and physical effort is possible. Noradrenaline is particulary sensitive to physical activity, while an increase in Adrenaline levels was shown to be more influenced by mental effort (see Wilson & Eggemeier, 1991). A NA/A ratio of 5 and higher is said to reflect physical activities, while a low ratio, between two and three, reflects mental effort (Fibiger et al., 1986). Recently, however, it was found that emotional stress, e.g. due to driving in heavy fog, can increase NA excretion (Vivoli et al., 1993, Van der Beek et al., 1995). Unpleasant, low-control tasks (e.g., vigilance tasks) have also been linked to raised Cortisol excretion, while high control tasks that require effort, were connected to increased Adrenaline and NA levels (see Raggatt & Morrissey, submitted).

Relating hormone levels to specific events is difficult, but as an index of health-threatening longer-term effects of stress, they have been shown to be very useful (e.g., Mulders et al., 1982, 1988, Raggatt & Morrissey, submitted).

**Event Related Potentials**

Compared to background EEG, certain low-amplitude potentials can indicate task demands. Most research has taken place regarding the amplitude and latency of positive potentials that occur minimally 300 ms after stimulus presentation, the $P_{300}$. Amplitude of the $P_{300}$-family increases with unexpected, task-relevant stimuli, and its latency parallels cognitive-evaluation time and increases with task
complexity (e.g., Brookhuis, 1989). \(P_{300}\) amplitude is an index of the subjects’ perceptual/central processing load, until the moment performance declines, then the amplitude remains unaffected (Gopher & Donchin, 1986). The amplitude also indexes the amount of resources allocated to a secondary task. In a primary-task-only-condition the \(P_{300}\) amplitude increases with task complexity. If the \(P_{300}\) is secondary-task-elicited it decreases with primary-task complexity increase (see Kramer, 1991, Humphrey & Kramer, 1994). In most studies a secondary-task technique is used in which a memory set has to be evaluated against stimuli. Only stimuli that are in the memory set elicit a \(P_{300}\). The use of the secondary-task technique in which subjects should not respond to frequent stimuli, but only to certain rare stimuli (‘Oddball Paradigm’) has the same disadvantages as any other secondary-task technique. Problems with artifacts, which can easily appear in low-amplitude physiological signals, can be added to these secondary-task-disadvantages. The main advantage of the ERP-technique is its high diagnosticity to perceptual/cognitive processing, and its insensitivity to response factors.

A relatively new technique is the irrelevant-probe method (see, e.g., Bauer et al., 1987, Hedman & Sirevaag, 1991, Sirevaag et al., 1993). This technique is low on primary-task intrusion and no overt responses to stimuli are required. The irrelevant-probe method uses as stimuli tones that are presented to the operator. The operator is instructed to ignore these tones. \(P_{300s}\) that are elicited by the irrelevant tones vary with primary-task workload in the same way as traditional secondary-task \(P_{300s}\). Again ERP amplitude decreases with increased task difficulty. In a rotary-wing-aircraft simulator study, Sirevaag et al. (1993) used this technique and found larger \(P_{300}\) amplitudes in a low-load condition than in a high-load condition. The authors conclude that in the low-load condition pilots apparently had sufficient capacity to process the irrelevant probes, while the demands of the high-load conditions precluded active processing. Low-probability probes (rare tones) resulted in larger ERP differences between conditions than high-probability probes (frequent tones).

The main problem of all ERP techniques is the poor signal-to-noise ratio. Though repeated stimulus presentation and signal averaging is no longer a prerequisite due to new equipment and single-trial techniques, ERPs are easily contaminated by other electrical signals (generated by the heart, eyes and muscles, or external sources such as 50 Hz power disturbance). An additional problem is the morphological characteristics of ERP waves that are subject to intra-individual variability (Humphrey & Kramer, 1994). Nevertheless, Humphrey and Kramer (1994) consider ERPs, in particular the \(P_{300}\), candidates for the assessment of dynamic changes in mental workload.
Electromyogram, the measurement of the electrical activity of task-irrelevant muscles (ElectroMyoGram, EMG) was previously directed towards limb-muscle activity, but is nowadays concentrated on the activity of facial muscles. Muscles are called task irrelevant if their activity is not required, either directly or indirectly, for the motor performance of a task. The origin of ‘task irrelevant’ activity of facial muscles lies in the medial interneurons in the lower pontine and medullary reticular formation that receive projections from the limbic system (Van Boxtel & Jessurun, 1993). The medial component would have a diffuse effect on the excitability of the motorneurons throughout the brain stem and spinal cord. Somatic and limbic influences converging on interneurons in the reticular formation could thus form the basis of nonvolitional, spontaneous activity of the facial musculature. This spontaneous activity has been defined as irrelevant activity. Differences between different facial muscles may be related to histochemical and physiological properties (see Van Boxtel & Jessurun, 1993). Facial muscles are strongly involved in expressive behaviour in social and non-social situations and these muscles have motor functions (e.g., the zygomatic muscle elevates the cheek to a smile), and may also function in the regulation of cerebral blood flow and temperature.

Van Boxtel & Jessurun (1993) reported that tonic activity of the following facial muscles reflects mental effort (see Fridlund & Cacioppo, 1986, for guidelines for electrode placement and EMG research): the lateral frontalis muscle, the corrugator supercilii and orbicularis oris inferior (see also Waterink & Van Boxtel, 1994). Activity of these muscles is considered an index of mobilization of general, non-specific resources. Not, or less, sensitive to mental effort is activity of the orbicularis oculi, zygomaticus major and the temporalis muscles. It was found that activity of the orbicularis oculi and zygomaticus major “may be representative of situations where suboptimal performance can no longer be compensated for by the mobilization of additional resources, a situation Sanders (1983) calls stress” (Van Boxtel & Jessurun, 1993).

It should be noted that facial muscle activity has also been related to the experience of emotion. Activity of the corrugator muscle has been shown to be related to exposure to negative visual emotional stimuli (e.g., a slide of a snake), while positive stimuli (e.g., happy faces) elicited activity of the zygomaticus muscle (Dimberg, 1988, Dimberg & Thell, 1988) and of the orbicularis oculi (Jäncke, 1994). Jäncke (1994) found no effect of emotionally charged stimuli on activity of the frontalis muscle. Compared with the corrugator muscle, the frontalis muscle may for this reason be preferred for mental effort-assessment. If on the other hand emotional evaluation is of interest, measurement of activity of the corrugator muscle may be preferred.

The assessment of mental effort by facial muscle activity is a fairly recent development. The results recited above seem to indicate
that facial EMG provides promising measures in the field of mental workload.

4.4 Relation between measurement groups

Dissociation of measures

Not all measures are sensitive to workload in the same area of performance, and ‘dissociation’ between measures of different categories have been reported (e.g., Vidulich & Wickens, 1986, Yeh & Wickens, 1988, see also Eggemeier & Wilson, 1991). In general dissociation between self-reports and performance measures are reported, although a few authors have found a dissociation between self-reports and physiological parameters (e.g., Myrtek et al., 1994). Measures dissociate if they do not correspond to changes in workload, or if one measure indicates a decrease in workload while the other indicates an increase. Yeh & Wickens (1988) offer as explanation of these dissociations a differential sensitivity of different measures to particular sources. Performance is affected by amount of resources invested, by resource efficiency and by competition for a resource, while subjective workload perception is affected by amount of resources invested and by demands on working memory. Motivation, task difficulty and subjective criteria of performance all determine the amount of resources invested (Yeh & Wickens, 1988). Regarding dissociation of measures, Gopher & Braune (1984) even question the sense and use of (self-report) measures of workload that are only weakly related to - or do not correspond to - the actual behaviour of subjects. Later on, in the same manuscript, they take a less strict position towards self-report measures and value them as conscious experience of workload.

It is questionable whether there really is a problem of dissociation of measures, in particular if a measure seems insensitive. Not all behaviour has to become overt in reduced performance, and not all measures have to be strongly correlated. In multidimensional concepts - and mental workload is likely to be a concept with multiple dimensions (chapter 2) - disagreement between subjective and objective measures may provide more information than does agreement (Muckler & Seven, 1992). In self-reports of workload, judgements on these multiple dimensions are integrated, sometimes giving the impression of divergence. The effort concept is also of particular interest here because, as mentioned previously, performance can remain stable while physiology (or self-report measures) indicate increased effort. As claimed before, this increase in effort can be maintained for limited periods of time, but clearly has its costs. It is therefore too simplistic to state that no reduced performance is equal to no increase in workload. It is also somewhat surprising that Vidulich and Wickens (1986) state that self-reports of workload are insensitive in the case of automatic
information processing and that this is due to the restricted representation of these processes in consciousness. Finding no effect on self-reports should not be unexpected, since automatic processing hardly uses any resources and therefore does not lead to an increase in workload.

Demand, in particular in the A3 region (see figure 2), might cause a dissociation between performance and the other measures, whereas in the C region performance and self-report ratings may ‘dissociate’. A good agreement between performance results and self-reports (Vidulich & Wickens, 1986) is only to be expected if performance is in the D, A2 and B region, and not in the A1, A3 or C regions.

Two groups of techniques

Gopher and Donchin (1986) argue that there are two groups of techniques to measure mental workload. The first group assumes that it is possible to obtain a global measure of mental workload, more or less comparable to single-resource use. Amongst these techniques are self-report measures, performance measures and physiological measures that are arousal-related. The other group of measures are procedures that are diagnostic, and are linked to theories of multiple resources. Secondary-task techniques and some of the physiological measures belong in this group. It is possible that single-resource theories and global workload measures are in many cases applicable, simply because task demands in one dimension predominate. Also, integration of dimensions is possible. In particular, self-reports and physiological measures that indicate a general arousal level could reflect integrated workload over different dimensions. Only when demands on certain dimensions are expected to be high, is there reason for apriori preference for measures from the diagnostic group. In general, and in particular in most applied settings, measures from both groups are useful.

Workload redline

If the workload redline is not determined by the point at which performance measures start to deteriorate (as was proposed by Rueb et al., 1992), but is determined by the point at which region A2 is departed, then performance measures alone are by definition not sufficient to determine whether load is unacceptable. Nevertheless, performance measures remain indispensable in redline research to determine whether workload is in the A region. Again, this is an argument in favour of the use of measures from multiple measurement

---

5 i.e., according to Gopher & Donchin, some techniques do claim to cover multiple dimensions separately

One of the aspects of workload measures that is emphasised in workload redline is the use of absolute versus relative measures. Traditionally, relative measures have been used. With relative measures, task performance, self-reports and physiological measures during baseline performance are compared with the same measures during performance of the task or system under evaluation. Some authors claim that absolute measures are required for workload redline (e.g., Wierwille & Eggemeier, 1993). So far, critical values on the SWAT rating-scale have only been proposed by Reid & Colle (1988). However, the critical SWAT value of 40 they mention refers to the point at which performance begins to be affected (the transition from region A to B). Such a workload redline is a primary-task workload margin (e.g., Wickens, 1984). This margin is defined as a critical level at which the (primary) task has to be performed. Beyond that point, primary-task performance is affected. Although performance margins can be successfully determined, an absolute criterion for workload itself, i.e. the critical value of a measure denoting that region A2 has been left, is in my view not tenable. The reason for this is that workload is a relative measure; it is the proportion of the capacity that is allocated for task performance. The amount of resources allocated does not only depend upon task demands, but also depends on capability or willingness to handle the demand. The conceptual problems of a workload redline become very prominent in applied settings. In traffic, for instance, the capabilities of individuals in the driving population vary to a great extent. Novice drivers have to allocate more resources for task performance than experienced drivers. Similar differences in capability exist between young and elderly drivers. Consequently, for the same task each individual has his or her own workload redline.

In spite of the problems associated with redline definition, an approach that includes primary-task performance margins relating to the cost of maintaining performance, is useful in any applied field of workload assessment. Self-report scales and performance measures (for the A to B region shift) are probably the most promising measure groups for this. Physiological indices that are opposed to baseline measurements can be very useful to assess operator effort; the cost of performance maintenance.

**Workload peaks**

Another source of ‘dissociation’ of measures could be workload peaks of relatively short duration. In most tasks the demands are not continuously at the same level, but differ over time. Measures of workload, however, are frequently aggregated over time. Over a complete period only one rating of the amount of effort that has been exerted is asked, and heart rate variability is calculated over periods of
30 seconds up to minutes. Performance-measures are also aggregated over time. There are only a few measures that can be directly related to workload peaks, e.g. the ERP measures that are related to a single stimulus event. If aggregated measures are taken in task situations where peak loads occur, caution is required. It is difficult to say in advance which aspect was rated by the operator in the self-report: overall workload or peak loads. Performance and physiological measures may or may not be sensitive to peak loads.

Verwey & Veltman (1995) have compared different measures’ sensitivity to peak loads in a driving task. In principle, driving is a suitable task for peak-load research, in particular because the road and traffic environment is continuously changing. In order to control the task demands, Verwey & Veltman made use of an artificial secondary task. A supplementary auditory or visual task was added to this to effectuate peak loads of 10, 30 or 60 seconds. All measures were analysed during or directly after peak loads, so no conclusion with respect to measure sensitivity to overall versus aggregated effects of workload peaks can be drawn from this study. Although the tasks that had to be performed were of a highly artificial nature and the ecological validity of this study is questionable, its merits are that attention is drawn to the largely neglected aspect of peak loads. Some of the results of the study will be discussed in the next chapter.