The Measurement of Drivers' Mental Workload

Dick de Waard

Published by The Traffic Research Centre VSC, University of Groningen, P.O. Box 69, 9750 AB HAREN, The Netherlands.

Printed by Drukkerij Haasbeek, Alphen a/d Rijn.

CIP-gegevens Koninklijke Bibliotheek, Den Haag

The measurement of drivers' mental workload / Dick de Waard. Haren: Verkeerskundig Studiecentrum, Rijksuniversiteit Groningen. -Ill. Proefschrift Rijksuniversiteit Groningen. -Met lit. opg. -Met samenvatting in het Nederlands
ISBN 90-6807-308-7
Trefw: mentale belasting, psychologisch onderzoek, verkeer

RIJKSUNIVERSITEIT GRONINGEN

THE MEASUREMENT OF DRIVERS' MENTAL WORKLOAD

Proefschrift

ter verkrijging van het doctoraat in de
Psychologische, Pedagogische en Sociologische Wetenschappen
aan de Rijksuniversiteit Groningen

op gezag van
Rector Magnificus, prof. dr. F. van der Woude,
in het openbaar te verdedigen op
donderdag 6 juni 1996
des namiddags te 14:45 uur (precies)

door

Dick de Waard
geboren op 17 juli 1964
te Enschede

Promotores:

Prof. Dr. J.A. Rothengatter
Prof. Dr. T.F. Meijman

Co-promotor:

Dr. K.A. Brookhuis

Aan mijn ouders

Dankwoord

Dit proefschrift is gebaseerd op onderzoek dat ik samen met anderen bij het Verkeerskundig Studiecentrum heb verricht. De uitvoering van het onderzoek, alsmede het schrijven van dit proefschrift, zou zonder de hulp, steun en opmerkingen van veel mensen niet mogelijk zijn geweest. Ik zou deze mensen dan ook heel erg willen bedanken. Allereerst dank aan mijn promotores, prof. dr. Talib Rothengatter en prof. dr. Theo Meijman, die elkaar mijns inziens erg goed aanvulden in hun commentaar op eerdere versies van het proefschrift. In het bijzonder gaat mijn dank uit naar co-promotor Karel Brookhuis, die mij wegwijs heeft gemaakt in de wereld van de toegepaste psychologie. Hij is van begin af aan bij al het onderzoek betrokken geweest en heeft altijd vertrouwen in mij gehad. Ik waardeer dit zeer. Peter Albronda wil ik ook graag met name noemen. Hij heeft telkens weer technische wonderen verricht, en zonder hem zou geen enkel van de in dit proefschrift beschreven onderzoeken kunnen zijn uitgevoerd. Prof. T.A.B. Snijders en Arie van Roon waren in een latere fase van het schrijven aan deze these bereid mij te woord te staan, en ik ben hen erg erkentelijk voor de geboden hulp. Margi Rothengatter dank ik ‘for improving my bad English’.

In de loop van het onderzoek, alsmede gedurende het schrijven van het proefschrift, zijn diverse mensen mij behulpzaam geweest. Graag bedank ik (in willekeurige volgorde): Stephen Fairclough (Thanks Stephen), Rob van Ouwerkerk, Frank Steyvers, Maaike Jessurun, Monique van der Hulst, Fokie Cnossen, Peter (Happy Boerenkool) Raggatt, Erik-Jan Westra, Adriaan Heino, Wiebo Brouwer, Frans Gort, Koen Kok, Harry Bakker, Joyce van Dorssen, Peter van Wolffelaar, Wim van Winsum, Erik Saaman, Peter Politiek, Jinke van der Laan, Ben Mulder, Leendert van der Linden, Wim Maring, Peter Lourens, Hans Oude Egberink, Gerbrand de Vries, Ton Rooijers en alle niet-genoemde (ex-) VSCers. Rineke Richters dank ik voor de regelmatig geboden hulp bij het opsporen van literatuur.

Natuurlijk ben ik Corine erg veel dank verschuldigd voor al haar steun, en Jolien, omdat zij mij altijd vrolijk stemt.

Table of contents

Summary 3
Samenvatting 7

1 Introduction 11

2 A model of mental workload, task performance and demands 21

3 Characteristics of measures 27

4 Measures 31
4.1 Self-report measures 31
4.2 Performance measures 34
4.3 Physiological measures 37
4.4 Relation between measurement groups 49

5 Traffic research 53
5.1 Self-report measures 58
5.2 Primary-task performance measures 64
5.3 Secondary-task performance measures 73
5.4 Physiological measures 79
5.5 Discussion 92

6 Conclusions related to driver workload measures 97

7 References 107

8 Appendices

1 Driving behaviour on a high accident rate motorway in the Netherlands
2 The effect of road layout and road environment on driving performance, drivers' physiology and road appreciation
3 The effects of mobile telephoning on driving performance
4 Elderly and young drivers' reactions to an in-car enforcement and tutoring system
5 Assessing driver status: a demonstration experiment on the road
A RSME, self-report effort scale
B Bartenwerfer's activation scale

Summary

Driving a vehicle may seem to be a fairly simple task. After some initial training many people are able to handle a car safely. Nevertheless, accidents do occur and the majority of these accidents can be attributed to human failure. At present there are factors that may even lead to increased human failure in traffic. Firstly, owing in part to increased welfare, the number of vehicles on the road is increasing. Increased road intensity leads to higher demands on the human information processing system and an increased likelihood of vehicles colliding. Secondly, people continue to drive well into old age. Elderly people suffer from specific problems in terms of divided attention performance, a task that is more and more required in traffic. One of the causes of these increased demands is the introduction of new technology into the vehicle. It began with a car radio, was followed by car-phones and route guidance systems, and will soon be followed by collision avoidance systems, intelligent cruise controls and so on. All these systems require drivers' attention to be divided between the system and the primary task of longitudinal and lateral vehicle control. Thirdly, drivers in a diminished state endanger safety on the road. Longer journeys are planned and night time driving increases for economic purposes and/or to avoid congestions. Driver fatigue is currently an important factor in accident causation. But not only lengthy driving affects driver state, a diminished driver state can also be the result of the use of alcohol or (medicinal) sedative drugs.

The above-mentioned examples have in common that in all cases driver workload is affected. An increase in traffic density increases the complexity of the driving task. Additional systems in the vehicle add to task complexity. A reduced driver state affects the ability to deal with these demands. How to assess this, i.e. how to assess driver mental workload is the main theme of this thesis.

In chapter 1, the theoretical aspects of mental workload are introduced. The difference between task demand, i.e. the external demand, the goals that have to be reached, and (work)load, i.e. the individual reaction to these demands, receive attention in this chapter. Mental workload is defined as a relative concept; it is the ratio of demand to allocated resources. Task difficulty is explicitly separated from task complexity. Task complexity would have been an objective property of the task that is related to demand on computational processes, were it not dependent upon individual goal setting. Task difficulty is very much dependent upon the context and the individual. Applied strategies may affect resource allocation or task complexity and thus difficulty and mental workload.

In chapter 2, a model of mental workload, task performance and demands is presented. In the model, performance and workload are related to task demands in 'regions of performance'. Two regions receive specific attention; namely those in which performance remains

unaffected at the cost of increased effort. A division between state-related effort and task-related effort is made. State-related effort is exerted in the case that the operator's state deteriorates but performance remains unaffected, while task-related effort is exerted to maintain performance in the case of increased task complexity. It is argued that both processes indicate increases in mental workload. Here, a key question arises: is it possible that different measures are differentially sensitive to these two kinds of effort?

In chapter 3, an overview of general characteristics of measures is given, while in chapter 4, specific measures of mental workload are presented. The often-used division between self-report measures, measures of task-performance and physiological measures is conserved. Different measures are presented and evaluated on their potential use as indicator of workload in traffic research. The issue of so-called dissociation of measures is weakened by the effort principle. The determination of a critical level of unacceptably high workload, the workload redline, is discussed and it is concluded that determination of a general valid level in terms of absolute values or scores on a measure is unattainable owing to individual differences in workload and the relativeness of the concept of mental workload. Performance margins are considered to be more useful in workload research than a workload redline.

In chapter 5, seven studies in which mental workload differed between conditions are presented. The studies are divided into two groups; studies in which the driver's state was affected and studies that included an increase in task complexity. The latter group was further subdivided into studies that involved an increase in complexity of the environment as opposed to studies in which a task was added. Two self-report scales, two primary-task performance parameters and three ECG-parameters as physiological measures were selected to assess sensitivity to mental workload. Differential measure sensitivity to mental workload associated with non-optimal driver state opposed to mental workload caused by increased task complexity receive specific attention. It seems that the evaluated subjective effort scale is sensitive to both kinds of effort, while the 0.10 Hz component of heart rate variability is more sensitive to task-related effort than to state-related effort. Task conception, task interpretation in terms of goal setting, is an important factor for the primary task measures. The need to perform at an optimal level on the primary task of lane keeping is absent, and most people allow for inaccuracies in steering. As a result, under some of the 'load-conditions' improvement in primary task performance measures was found. This unexpected effect may be related to increased effort.

In chapter 6, the conclusions are summarized and the different measures are linked to the model of mental workload, task performance and task demands. Recommendations for the measurement of mental workload in applied settings are given in this chapter and the different

concepts that are related to mental workload are evaluated on the basis of the results of applied (traffic) psychological experiments.

Finally, as appendix, detailed reports on five of the seven experimental studies are included.

Het meten van de mentale belasting bij bestuurders

Autorijden lijkt een simpele taak. Na enige initiële training zijn veel mensen in staat om veilig een voertuig te besturen. Niettemin gebeuren er veel ongevallen en de meerderheid van deze ongevallen kan worden geweten aan menselijk falen. Heden ten dage zijn er verschillende factoren aan te wijzen die ertoe kunnen leiden dat dit falen in het verkeer zelfs zal toenemen. Allereerst is er, onder andere als gevolg van de toegenomen welvaart, de groei in het wegverkeer. Een toename in verkeersintensiteit betekent dat er hogere eisen worden gesteld aan het menselijke informatie verwerkingssysteem en dat de kans op botsingen tussen verkeersdeelnemers toeneemt. Op de tweede plaats is het zo dat mensen tot op hoge leeftijd blijven autorijden. Oudere mensen hebben specifieke problemen, met name problemen met het verdelen van de aandacht, en het is nu juist dit vermogen dat steeds vaker vereist is in het verkeer. Een van de oorzaken hiervan is de introductie van nieuwe technologie in de auto. Dit proces begon met de autoradio, werd gevolgd door autotelefoons en route-geleidingssystemen en zal spoedig worden gevolgd door anti-botssystemen, intelligente cruise-controls enzovoorts. Al deze systemen eisen aandacht op, aandacht die verdeeld moet worden tussen het systeem en de primaire taak van longitudinale en laterale voertuigcontrole. Op de derde plaats vormen bestuurders in een verslechterde gesteldheid een bedreiging voor de verkeersveiligheid. Steeds langere ritten worden gepland en het ’s nachts rijden neemt toe ten behoeve van economische doelen en/of om files te vermijden. Vermoeidheid van de bestuurder is een belangrijke ongevals-veroorzakende factor. Maar niet alleen lange ritten hebben een invloed op de gesteldheid van de bestuurder, een verslechterde toestand kan ook het gevolg zijn van het gebruik van alcohol, sederende medicijnen of drugs.
Bovengenoemde voorbeelden hebben gemeen dat in alle gevallen de mentale belasting van de bestuurder toegenomen is. Zo verhoogt de toenemende verkeersintensiteit de complexiteit van de rijtaak. De taakcomplexiteit wordt ook groter door extra apparatuur in de auto terwijl een verslechterde toestand van de bestuurder het kunnen omgaan met deze taakvereisten vermindert. Het centrale thema in dit proefschrift is hoe je veranderingen in de mentale belasting van bestuurders, zoals in bovengenoemde gevallen, meet.
In hoofdstuk 1 worden de theoretische aspecten van mentale belasting geïntroduceerd. Het verschil tussen taakvereisten, de externe eisen of de doelen die bereikt dienen te worden, en mentale belasting, de individuele reactie op deze taakvereisten, wordt in dit hoofdstuk belicht. Mentale belasting wordt gedefinieerd als een relatief concept, het is de verhouding tussen taakvereisten en beschikbaar gestelde hulpbronnen

(‘resources’). Taakmoeilijkheid wordt expliciet gescheiden van taakcomplexiteit. Taakcomplexiteit zou een objectieve eigenschap van de taak zijn die gerelateerd is aan computationele processen, ware het niet dat taakcomplexiteit afhankelijk is van individueel gestelde doelen. Taakmoeilijkheid is sterk afhankelijk van context en individu. Toegepaste strategieën kunnen invloed hebben op de toedeling van hulpbronnen of op de taakcomplexiteit, en dus op de moeilijkheid en de mentale belasting.
In hoofdstuk 2 wordt een model van mentale belasting, taakverrichting en taakvereisten gepresenteerd. In het model worden ‘prestatie regio’s’ beschreven waarbij taakverrichting en mentale belasting aan taak-vereisten worden gerelateerd. Twee regio’s krijgen in het bijzonder aandacht, in deze regio’s blijft het niveau van taakverrichting onaangetast ten koste van toegenomen inspanning (effort). Er wordt hierbij een onderscheid gemaakt tussen toestand-gerelateerde inspanning en taak-gerelateerde inspanning. Toestand-gerelateerde inspanning wordt geleverd om het taakniveau gelijk te houden terwijl de toestand van de taakverrichter verslechterd. Taak-gerelateerde inspanning wordt geleverd om het niveau van taakverrichting gelijk te houden in geval van toegenomen taakcomplexiteit. Er wordt gesteld dat beide processen een indicatie van verhoogde mentale belasting zijn. Hier komt tevens een van de sleutelvragen naar voren; ‘is het mogelijk dat verschillende mentale belasting-maten op verschillende wijze gevoelig zijn voor deze twee vormen van inspanning?’.
In hoofdstuk 3 wordt een overzicht van de algemene karakteristieken van maten voor mentale belasting gegeven, terwijl in hoofdstuk 4 de specifieke maten worden besproken. De veel-gebruikte verdeling tussen zelf-rapportage-, verrichtings- en fysiologische maten van mentale belasting wordt daarbij aangehouden. Verschillende maten worden besproken en geëvalueerd op hun potentiële nut als maat voor de indicatie van mentale belasting in verkeersonderzoek. De kwestie van de zogenaamde dissociatie van maten wordt afgezwakt met behulp van het inspannings- (effort) principe. Tevens wordt het bepalen van een kritiek niveau van onacceptabele hoge mentale belasting, de ‘workload redline’, besproken. Er wordt geconcludeerd dat het bepalen van een algemeen geldig niveau in termen van absolute waarden of scores niet haalbaar is vanwege individuele verschillen in mentale belasting en het concept mentale belasting, wat een verhoudingsbegrip is. In plaats van een ‘workload redline’, worden taakverrichtingsmarges als zijnde nuttiger beschouwd in het onderzoek naar mentale belasting.
In hoofdstuk 5 worden kort zeven experimenten beschreven waarin de mentale belasting tussen de condities verschilde. De onderzoeken worden ingedeeld in twee groepen; onderzoek waarin de toestand van de bestuurder verminderd is en onderzoek waarbij de complexiteit van de uit te voeren taak toeneemt. Deze laatste groep wordt verder onderverdeeld in onderzoek waarbij de toename in complexiteit in de taakomgeving ligt en onderzoek waarbij een extra taak wordt

toegevoegd. Twee zelf-rapportage schalen, twee primaire taakverrichtingsparameters en drie ECG parameters als fysiologische maten worden geselecteerd om vast te stellen hoe gevoelig deze maten zijn voor mentale belasting. Hierbij wordt speciale aandacht geschonken aan de differentiële gevoeligheid van maten voor mentale belasting in geval van een niet-optimale toestand van de bestuurder versus mentale belasting als gevolg van toegenomen taakcomplexiteit. Het lijkt erop dat de geëvalueerde subjectieve inspanningsschaal gevoelig is voor beide soorten van mentale inspanning, terwijl de 0.10 Hz component van de hartslagvariabiliteit gevoeliger is voor taak-gerelateerde inspanning dan voor toestand-gerelateerde inspanning. Interpretatie van de taak in termen van het stellen van doelen is een belangrijke factor voor de primaire taakverrichtingsmaten. De noodzaak om op het hoogste niveau te presteren bij de primaire taak van ‘het tussen de lijnen houden van het voertuig’ is afwezig, en de meeste mensen staan derhalve onnauwkeurigheden in de stuurcorrecties toe. Dit heeft tot gevolg dat in sommige van de ‘verhoogde mentale belasting condities’ de primaire taakverrichting verbetert. Dit onverwachte effect is mogelijk gerelateerd aan toegenomen inspanning.

In hoofdstuk 6 worden de conclusies op een rij gezet en worden de verschillende maten in verband gebracht met het model van mentale belasting, taakverrichting en taakvereisten. Ook worden aanbevelingen gegeven voor het meten van mentale belasting in toegepaste settings en worden de verschillende concepten die gerelateerd zijn aan mentale belasting geëvalueerd op basis van de resultaten van toegepaste verkeerspsychologische experimenten.

Tenslotte is als appendix een gedetailleerde rapportage van vijf van de zeven experimenten toegevoegd.

1
Introduction

Over the past thirty years, the difficult tasks that operators, in particular aircraft pilots and air traffic control operators, have had to perform have drawn attention to the area of mental workload. General questions have been asked such as "How busy is the operator?", "How many tasks can he handle safely?" and "Does the operator have to ‘try hard’ to maintain an adequate level of performance?". If task demands are high in relation to the operator’s capabilities, errors may occur, and in interaction with neglected classical human factors issues such as a proper layout of instrumentation panels, these errors may become critical for safety. Even economic interest can raise workload-related questions. As an example, Wickens (1992) described a controversy between an airline industry and a pilot association. The airline industry claimed that a certain class of aeroplanes could be flown by two crew members, while the pilot organization claimed that demands at peak times would be excessive and would require a three person complement. Such issues have called for a definition of mental workload and the methods to assess it.

This thesis is about how to measure driver workload. In the present chapter the more theoretical aspects of mental workload in general and driver mental workload in particular will be introduced. In chapter 2 a model that relates task demands to workload and performance will be presented. In chapter 3 and 4, the general criteria for workload measurement techniques are described, followed by a categorization of measures. Properties of different measurement techniques and experience from non-traffic research will be reported in chapter 4. From chapter 5 onwards the focus is on the use of the techniques in traffic research. Although some of the techniques have been applied in traffic research, an overview and review of their characteristics in this specific field is missing. Driving is a very dynamic task in a changing environment. Moreover, contrary to many laboratory tasks, the driving task is to a large extent influenced by drivers themselves. The driver’s influence on the task ranges from strategic aspects such as route selection, to ‘control behaviour’ such as the accuracy in lane keeping. In particular, the increase in RTI (Road Transport Informatics) makes the evaluation of mental workload techniques for use in traffic research relevant and urgent. With an increase in in-vehicle RTI applications, road safety may be negatively affected. Much is to be gained by thorough evaluation of the mental load effects of new equipment before introduction to the market (e.g., Parkes, 1991). In chapter 5, the measurement techniques will be evaluated on sensitivity, reliability and operational aspects on the basis of results of several field studies. Sensitivity and dissociation of different measures will also be evaluated in the context of the mental workload model presented in chapter 2. The so-called ‘workload

redline', which indicates the critical level of too much mental workload, will be linked to the model and its potential, as well as the problems associated with correct redline determination, will be discussed.

Theories relevant for mental workload

In order to understand mental workload, the introduction of some basic concepts is required. The concept of a limited processing *capacity* can be found in many theories (e.g., Broadbent 1958, Kahneman, 1973, Posner, 1978, Wickens, 1984). Kahneman (1973) specifies the metaphor of a single undifferentiated capacity (the 'modal' view) from which *resources* are available for task performance. O'Donnell & Eggemeier (1986) make no difference between the metaphoric words 'capacity' and 'resource' and use the words as interchangeable terms. Wickens (1992) disagrees with this. He defines capacity as the maximum or upper limit of processing capability, while resources represent the mental effort supplied to improve processing efficiency. This is in line with Norman & Bobrow (1975) who also refer to resources as processing effort. In this thesis the differentiation between capacity as upper limit of capability and resources as amount of processing facilities allocated will be followed. Resources are characterized by two general properties: their deployment is under voluntary control and they are scarce. Only the very simple resource models consider capacity to be fixed. According to Kahneman (1973) there is some elasticity in capacity and the availability of resources, the mobilization of resources could be increasingly possible, e.g. as a result of increased processing load.

The relation between resource allocation and task performance is supposed to be linear, until the moment all resources are invested. From that point on, no more resources can be invested and task performance will remain stable. Norman & Bobrow (1975) call such a task resource-limited. The resource-limited task is opposed to a data-limited task. When performing a data-limited task, additional available resource investment does not lead to increased performance due to limitations in data quality. Although the theory could be applied to a variety of situations, it could not explain why effective time-sharing and unaffected performance could occur when a second auditory task was added to a primary visual task.

In the 1980's Wickens proposed a multiple-resource theory in which different resources for different modalities are assumed (Wickens, 1984). Most prominent are the auditory and visual resources. In addition to these, central resources are supposed, which are required for the performance of almost all tasks. An overlap in resource requirement, e.g. the performance of two auditory tasks, soon requires full auditory capacity use. In that case performance on both tasks will be affected. Tasks that require different resources, e.g., a visual task combined with an auditory task, will not directly interfere with each

other and performance of either task can remain unaffected, provided there is no performance decrement caused by central resource use.

The concept of multiple resources is connected to three dimensions. The first dimension is the processing *stage*, i.e. perception (including encoding), central, and response processing. The second dimension is *modality* of input and response. The auditory, visual and tactile modality draw upon different resources and cross-modal time-sharing can be better performed than intramodal timesharing. Listening to someone and watching something at the same time associate better than listening to two things at the same time. The third dimension is the processing *code*. The processing code can be either verbal or spatial.

With respect to stage, the multiple-resource theory predicts more interference between tasks if both tasks demand spatial processes, or if both demand verbal processing *across any stage*. So, even if the perceptual modality is different (e.g. auditory and visual) the tasks will interfere if both require (e.g.) verbal central processing. The second dimension, separateness of modality resources, was later dropped by Wickens (1991), mainly due to the influence of physical restrictions. Two competing visual channels cannot be watched at the same time and hence require scanning, an additional cost. Moreover, two simultaneously presented auditory messages will mask one another. In the last dimension, different codes can be better combined. A manual, spatial, process can, for instance, be successfully time-shared with a visual process. A well known example is typing and sight-reading.

Capacity theories have been linked to *computational processes* and to *energetical mechanisms* (G.Mulder, 1986). In the processing of information from information uptake to overt or covert reaction, a series of stages are passed in which computational processes are performed. At least four stages of processing are identified (e.g., Sanders, 1983); stimulus preprocessing, feature extraction, response choice and response adjustment. Each stage is related to a processing module with a limited capacity. A large number of these processes are not conscious. These processes are fast and automatic and cannot be subjectively assessed (Meijman & Mulder, 1992). There are, however, other processes that require working memory and are (partly) conscious and can be subjectively determined. These two classes of processes have often been labelled automatic resp. controlled processing (see below). Electrical and magnetical brain activity during the performance of information processing tasks can help to identify which brain mechanisms are mobilised in different stages of information processing (e.g., Brookhuis, 1989, Wijers, 1989).

Energetical mechanisms facilitate the availability of computational processes, and depend upon the mental or physical state of the individual. Three energetical resources have been identified (Pribram & McGuiness, 1975); arousal, activation and a compensatory resource labelled ‘effort’. Note that *the resource* is labelled effort; this

should not be confused with the allocation of resources that Norman & Bobrow (1975) indicated as processing effort. The effort mechanism is active in the case of attention demanding information processing, or in the case that the operator's state differs too much from the required state. This last condition has been put forward by Hockey in his State Control Theory (Hockey, 1986). According to this theory, central executive mechanisms compare the current cognitive state with a required or target state. Whenever there is a mismatch between these two states the energetical construct of effort can be involved in actively manipulating the current state towards the target state. Hockey calls such a manipulation 'state management'. By investing mental effort the detrimental influences of stressors (such as noise, information overload or monotony) can be successfully counteracted. A task of a highly monotonous nature, for instance, may stimulate compensatory mental effort to maintain performance. In his state-control theory, Hockey (1986) also puts forward the aspect of strategy. A minimal strategy for example is one of inaction. Performance will probably not be very high, while the effort costs are always low. Another option is that, instead of adapting the current state, different criteria for optimal performance are accepted. However, this type of goal changes often result in decreased performance. A last option is to deal directly with the source of environmental influence. A window can be opened in order to regulate environmental temperature, or it can be closed in order to reduce the noise level (Van Ouwerkerk et al., 1994a).

Sanders (1983) and G.Mulder (1986) have put forward an integration of energetic and computational models. In this model the efficiency of computational processes is affected by the energetical resources: arousal, activation and effort. Arousal affects feature extraction, while activation affects the motor organization. In tasks that require retention of information in the Working Memory, the effort mechanism is the structure that supplies energy for processing. G.Mulder (1986) assumes that there are two forms of effort: effort for tasks that require controlled information processing (computational effort), and effort in the case that an individual has to change the current energetical resource state towards a required state (compensatory effort). Cnossen (1994) labels the first as *task-related effort* and the latter as *state-related effort*.

In the information processing and task performance literature two types of theories dominate; physiological theories and cognitive theories. Quite often adherents of these two theories make use of the same terminology, which very much complicates understanding. Sometimes resources are referred to as processing modules with a limited capacity, while at other times, resources are referred to as physiological energetical structures. Nevertheless, Sanders (1983) and G.Mulder (1980, 1986) have made clear that these two types of theories are not mutually exclusive, and have proposed an integration. In the

following paragraph the concepts that are important for the measurement of mental workload will be described. Links to both cognitive and physiological theories remain apparent.

The concept of mental workload and its assessment

A simplistic definition of workload is that it is a demand placed upon humans. This definition attributes workload exclusively to an external source. An indication of workload, however, can be better defined in terms of experienced load. With experienced load, workload is not only task-specific, it is also person-specific (Rouse et al., 1993). Not only individual capabilities, but also motivation to perform a task, strategies applied in task performance, as well as mood and operator state, affect experienced load. In the (mental) workload literature, task demands and the effect of these demands on the operator are sometimes, unfortunately, indicated with the same term, 'workload'. For reasons of clarity in this text *demand* will henceforth be used to indicate the task demands. Demand is determined by the goal that has to be attained by means of task performance, and is, once the goal has been set, external and independent of the individual. *Load* or *workload* will be used to describe the effect the demand has on the operator in terms of stages that are used in information processing and their energetics. More specifically, workload is the specification of the amount of information processing capacity that is used for task performance. In the concept of mental workload how the goal is reached (e.g. the order of actions) and individual restrictions imposed upon performance (e.g. in terms of accuracy or speed) are included. Therefore workload depends upon the individual, and owing to the interaction between operator and task structure, the same task demands do not result in an equal level of workload for all individuals. Directly related to demand is (task) *complexity*. Complexity increases with an increase in the number of stages of processing that are required to perform a task. Task demand and complexity are mainly external, but both depend upon (subjective) goals set for task performance. *Difficulty* of a task is related to the processing effort (amount of resources) that is required by the individual for task performance, and is dependent upon context, state, capacity and strategy or policy of allocation of resources. Kantowitz (1987) has proposed this differentiation between complexity and difficulty as a property of, respectively, the task in isolation versus the interaction between task and individual. The parallel with, e.g., a maths exam is noticeable; the goal that has to be reached, solving the mathematical questions, is the same for everyone and depends upon the number of calculations that have to be performed. However, goal setting affects the task demands, and there is a difference between 'considering a C sufficient versus going for an A'. How difficult the calculations are depends very much upon the individual who has to perform the calculations. They may be relatively easy for a trained or experienced person and very hard for a novice. After a sleepless night,

however, the task will be more difficult even for the experienced person.

O'Donnell & Eggemeier (1986) define workload as that portion of the operator's limited capacity that is actually required to perform a particular task. Workload measurement is the specification of the amount of capacity used. In this definition also, workload is not solely task-centred. Mental workload depends upon the demands in relation to the amount of resources the operator is willing or able to allocate, and is therefore a relative concept (Meijman & O'Hanlon, 1984, Zijlstra & Mulder, 1989).

In workload measurement, not only processing effort or resource allocation (Norman & Bobrow, 1975) are of primary importance, the term *effort* is also used for the mobilisation of additional resources as a compensatory process (see G.Mulder, 1980, Aasman et al., 1987, Vicente et al., 1987). Effort reflects the operator's reaction to demand and the amount of effort being expended is considered by many to be one of the most important components of (if not equal to) mental workload. Vicente et al. (1987) mention two important reasons for this. Firstly, the effort expended by the operator is not necessarily related to input load (demand). The operator's reaction to the demand depends on internal goals and adopted criteria or strategies. Secondly, there is no simple relationship between performance and effort invested. The expended amount of effort depends very much on the structure of the task (data-limited versus resource-limited, Norman & Bobrow, 1975) and, related to this, the amount of practice and experience, and of the operator's state.

G.Mulder (1980) has linked mental workload to a 'controlled mode' of information processing. A distinction between two modes of information processing has been proposed (see Schneider & Shiffrin, 1977, Shiffrin & Schneider, 1977): automatic versus controlled information processing. Automatic processing is fast, not conscious, rigid, requires almost no resources or attention and can be performed in parallel. Automation follows frequent, consistent practice. Controlled processing is effortful, serial, conscious, and is flexible. Controlled processing requires the retention of information in working memory, and hence requires resources and attention. According to G.Mulder (1980) the amount of time an operator processes information in this controlled mode is a reflection of mental effort. Also, in general, a task with higher mental demands is expected to lead to a proportional increase in controlled processing time (see also Meijman & O'Hanlon, 1984).

Resource models have traditionally been used extensively in mental workload research (e.g., Gopher & Sanders, 1984) and the framework has proven to be useful in this area of research. However, this does not mean that the *multiple*-resource model is universally

supported. Kantowitz (1987), not an opponent of capacity theory, has criticized its multidimensionality. Kantowitz considers the theory "too powerful and too difficult to reject" and states "I do not trust a model that cannot be falsified" (p. 91). He suggests that it is too easy to add another resource ('pool of capacity') if data do not fit the theory and he draws attention to a hybrid model launched previously by himself and Knight (Kantowitz & Knight, 1976). In that model a single pool of capacity is divided between perceptual and response stages of information processing by a Static Capacity Allocator. However, this model does not pay attention to interference within versus interference between modalities, a very useful aspect of the multiple-resource theory in mental-workload research. Here the multiple-resource theory has functional utility in predicting interference between tasks.

Clearly, the assessment of workload is coupled with task difficulty as experienced by the operator (Gopher & Donchin, 1986), in particular because several reactions to the task demands are possible. Operators can adapt their behaviour and cope with an increase in demand. They can also change their strategy and task goals and accept a lower performance level or they can give up completely (see Meijman & O'Hanlon, 1984). Strategies will also differ between individuals, and some strategies will be more effective and require less effort to reach the same level of performance. In the case of coping with the demand, an increase in effort is exerted while performance remains at the same level. In that case, performance measures will not reflect any change and be insensitive to the increase in workload, while other measures, such as self-report ratings or physiological measures, may well give an indication of effort exerted. In other conditions in which a change in strategy or 'quitting' behaviour occurs, measures of effort may remain unchanged or even show a decrease, while performance measures will indicate decreased task performance.

Terminology in mental workload research has its roots in cognitive and physiological theories. As a result, the terms used are sometimes unclear, as different authors use the same terms with differing meanings. In this thesis task demands, workload and effort are prime concepts. Task demands are determined by goals that have to be reached by performance. These goals can be defined in general terms such as 'the aircraft should land safely'. It is important to acknowledge that sub-goals are quite often self-set, e.g., first action A then B (or the other way around), and that giving priority to sub-goals can influence general goals and demand. Workload is the result of reaction to demand; it is the proportion of the capacity that is allocated for task performance. Effort is a voluntary mobilisation process of resources. State-related effort is exerted to maintain an optimal state for task performance while task-related effort is exerted in the case of controlled information processing.

Driver workload

A model of the main task of the driver is useful in mental workload research in driving. Parkes (1991) defines the primary task of the driver as "safe control of the vehicle within the traffic environment". As stated in the introduction, car driving is a dynamic control activity in a continuously changing environment. The driving task is not only influenced by the drivers themselves, but also by the behaviour of other traffic participants. It is not an easy task to model driver behaviour. However, a useful model of driving that takes more into account than just 'safe control' (Parkes, 1991) has been offered by Michon (1971, 1985) and Janssen (1979). In this model, car driving is described as a complex task with processes at a minimum of three hierarchical levels. At the top level, the *strategic level*, strategic decisions are made, such as the choice of means of transport, setting of a route goal, and route-choice while driving. At the intermediate level, *the manoeuvring level*, reactions to local situations including reactions to the behaviour of other traffic participants, take place. At the lowest level, *the control level*, the basic vehicle-control processes occur, such as lateral-position control. At this level automatic processes occur, while a level higher controlled processing is required. In particular driver-performance measures can be connected to the three levels. For example, steering-wheel movements reflect performance at the lowest level, car following performance and mirror looking are processes at the manoeuvring level, while errors in route choice reflect performance at the strategic level. Demands at all three levels can exceed capacity, and may result in affected performance, and that includes affected performance at other levels. A student driver cannot yet perform all control-level tasks automatically, and workload with respect to vehicle control is high. This may result in neglect of higher level tasks, such as mirror-checking. In a new traffic environment, e.g. driving in heavy traffic in a city abroad, manoeuvre-level tasks may put high demands on visual and central resources leading to affected performance on the other levels. Demands of monitoring other traffic could be so high that following the signs 'Antwerpen' is not possible and a turn is missed. In general performance at a higher level will be affected, although it cannot be excluded that under conditions of high manoeuvre-level demand some drivers will, e.g., shift to a wrong gear.

Sources of driver workload may be found both inside and outside the vehicle. A complex junction that has to be crossed or an important conversation on the car-phone will both increase task demands. Since driving is to a very large extent a visual task, demands on visual and central resources will be highest. However, in the years to come, more in-car technology will be installed in vehicles requiring a raise in allocation of auditory resources. The use of car-phones is already widespread and various new electronic intelligent in-car devices are being developed, the use of which will only increase. While it is

unlikely intentional, these devices will increase driver mental workload and possibly affect behaviour negatively, thus becoming a threat to traffic safety. There is another problem with this boost in development of equipment; collision-avoidance systems, traffic-information systems, driver impairment monitors and navigation systems individually can help drivers, but the combined use can result in overload of their information-processing system (Verwey, 1990). In the GIDS[1] project (Michon, 1993) this problem was recognized and the project proposed to add a scheduling system that plans information presentation (Verwey, 1993a). In scheduling, the driver's personal limitations should be taken into account. But before tasks can be properly scheduled, the effects on driver workload of tasks in isolation and the effects in combination with other tasks that require simultaneous performance, have to be assessed. A GIDS system needs information about the effect of each individual task on workload, preferably dependent upon local situations, before such a system can decide which task or signal to postpone. Likewise, a road authority might like to know whether the road layout at a specific accident blackspot increases driver mental workload before taking action, or a telecom company may decide to promote their voice-activated dialling car-phone that, in terms of workload, can be more safely combined with the primary task of car driving.

Under certain circumstances it is possible that new in-car technology will have the *opposite* effect of driver overload, and will lead to monotony in task performance. This could happen, as Kantowitz (1992a) pointed out, if new devices are to actually control the vehicle, similar to flight management systems in aviation (see also Wiener, 1987). At present, driver deactivating situations are mainly confined to monotonous motorway driving. The number of these low-stimulus conditions, in which the driver may become deactivated, may however increase if more functions are taken over by technology. There are scenarios for the future in which vehicle control in terms of steering-wheel movements will also be carried out by an automated system, and the "driver's" actions will be restricted to strategic level decisions (Hancock & Parasuraman, 1992).

A list of factors that affect driver workload is given in table 1. The table displays both driver state, trait and environmental factors that have an influence on workload. Factors may either increase or decrease mental workload. Automation and the allocation of functions may help the driver, e.g. in conditions where environmental demands are high, but could also turn driving into a task of vigilance. In general, feedback is intended to reduce demand, but sometimes it increases workload by providing additional information that has to be processed. High road-

[1] Generic Intelligent Driver Support

environment demands, e.g., having to merge in heavy traffic, increase workload, while the effects of alcohol, persisting monotony and fatigue increase workload by a reduction in capacity (Schneider et al., 1984, Kantowitz, 1992a, Wierwille & Eggemeier, 1993).

Table 1. Factors affecting workload.

Driver *State* Affecting Factors
monotony
fatigue
sedative drugs
alcohol
Driver *Trait* Factors
experience
age
strategy
Environmental Factors
road environment demands
traffic demands
vehicle ergonomics (RTI)
automation
feedback

2
A model of mental workload, task performance and demands

In chapter 1 the different concepts that play a role in driver mental workload were introduced and defined. The task that has to be performed or the goal that has to be reached can be described in objective terms. Goal setting (task conception or subjective task interpretation) determines the task goal that has to be reached (in terms of accuracy or speed) and thus affects task demand. Task demand can now be described in terms of operating stages which determine task complexity. How well the task is performed is an objective measure, namely the level of performance achieved. However, how the task is experienced, i.e. task difficulty, is not an objective property. Task difficulty depends upon task complexity, the operator's possibilities (i.e. capacity), his or her state and the applied strategy. Finally mental workload, the central concept in this thesis, is determined directly by task difficulty. On the basis of task difficulty processing resources are allocated and mental workload is reflected by the amount of allocated resources.

A relation between task demand and task performance has been described by Meister (1976, see also O'Donnell & Eggemeier (1986) for an adapted reprint). Meister defined three regions, region A, B and C. Region A is described as low operator workload with high performance. An increase in demands does not lead to performance decrements. In region B the level of performance declines with increased task demands. So, region B is the region where performance decreases with increases in demand, and increases in workload. In region C extreme levels of load have diminished performance to a minimum level, and performance remains at this minimum level with further increases in demand (see figure 1).

According to this model, a primary-task workload measure, i.e. a measure of performance, will only be sensitive to variations in levels of workload in region B. In region A performance remains stable and is independent of variations in demand, while in region C performance will remain at a minimum level, independent of demand. Other measures, e.g., self-report measures of workload, may be sensitive in region B and may clearly reveal overload in the C-region, while they need not to be sensitive in region A.

While extreme levels of load resulting in overload can be situated in the C-region, it is not clear where the domain of underload is. A relation between arousal, task difficulty and performance as was first found almost a century ago, the so-called 'inverted U', could help to complete the region model. In 1908 Yerkes & Dodson (Yerkes &

Dodson, 1908) published a famous paper on performance in a learning task under various levels of stress. The original paper did not describe the performance of human subjects, but of mice. Possibly due to their behaviour as a consequence of electric shocks that were administered, these mice were even called ‘dancers’. The major result of the study, which led to what later became the inverted-U hypothesis, was that

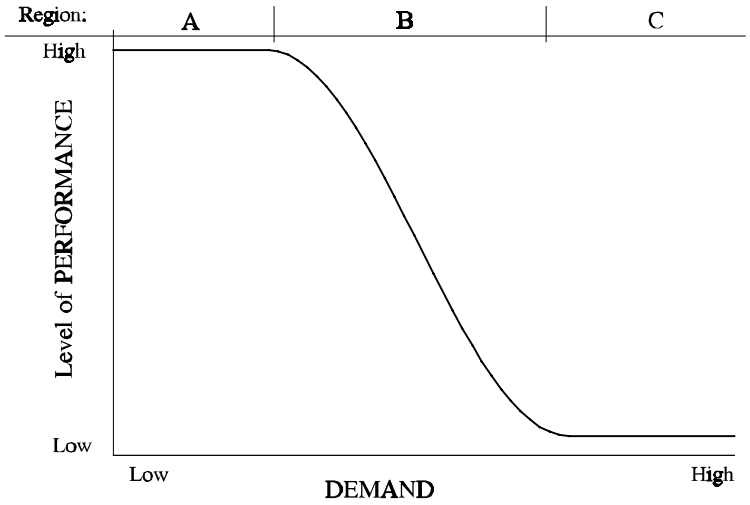


Figure 1. Hypothetical relationship between demand and performance (based on Meister, 1976)

with different strengths of stimulation the medium electrical stimuli were more favourable to the acquisition of a ‘habit’ than stronger or weaker stimuli. Although the original Yerkes-Dodson paper described a relation between stimulus strength and learning, the law has implicitly been broadened to account for the effects of arousal level on performance (Hebb, 1955, see Teigen, 1994, for a discussion of the law’s history). To return to the region model, this model could be completed by adding a deactivation or D-region at the far left end. The effects of monotonous tasks, for example, are situated in the D-region. These are low *demand* tasks that can result in *in*creases in task difficulty and workload by a reduction in capacity. In case of, e.g., boredom a reduction in capacity requires that a larger proportion of the capacity is used for performance of the same task, thus increasing mental workload (Meijman & O’Hanlon, 1984, O’Hanlon, 1981). It may also be that an affected state impedes the allocation of resources. By means of the addition of the D-region the complete inverted-U is split into four regions, the D, A, B and C regions.

A question that comes to mind with respect to the region model is "How much workload is too much?". This issue is usually

referred to as the determination of a *workload redline* (Reid & Colle, 1988, Wierwille & Eggemeier, 1993). When trying to tackle the determination of a redline there is a need to first decide upon the context of 'too much'. Degraded performance may indicate too much workload, but affected personal well-being is equally valid. Preliminary work on workload redline puts this line at the transition from region A to B (Rueb et al., 1992). Reid & Colle (1988) related just detectable performance decrements to self-report ratings, and this workload rating designated the absolute workload redline. The point of a just detectable performance decrement is at the transition from region A to B. While it is clear that performance measures themselves have defined the A-region, it may be useful to split the A-region up into three parts. In the middle part, region A2, the operator can easily cope with task demands and performance remains at a stable level with increases in demand without increased effort. In the A3 region, however, performance measures still do not show a decline, but the operator is only able to maintain the level of performance by increasing effort. Temporary compensation by the exertion of effort in region A3 is one of the advantages of human flexibility and is not critical. If, however, continuous effort is required to maintain performance, or if peak loads occur frequently, this can lead to stress, an unhealthy situation that has to be avoided (Zijlstra & Mulder, 1989, Meijman, 1989). This is in particular true if the operator has no control over the situation (e.g., Van Ouwerkerk et al., 1994b). It may therefore be more useful to put a workload redline at the transition from region A2 to A3 instead of at the transition from region A (A3) to B, as Rueb et al. (1992) did. In this way, the word *workload* redline remains related to *workload* instead of relating it to primary-task performance breakdown. A similar situation exists at the region that is to the right of the D-region, region A1. Here for instance monotony starts to affect the operator's state, but by 'trying harder', i.e. by the investment of effort, the primary-task performance level is not yet affected. A second workload redline then arises at the transition from region A2 to A1, where the operator is effectively counteracting a reduced operator state. When effort investment is no longer effective, the D-region is entered where performance is affected.

When demand increases, starting from the optimal operator state in region A2, the operator's capability of (effort) compensation will be exceeded at a certain moment and a transition from the A3 to the B region takes place. In the B-region performance is affected and at the moment that it has deteriorated to a minimum level the C region is entered. Task performance and workload as a function of demand are depicted in figure 2. It is important to stress that demand on the x-axis in figure 2 is not directly linked to region of performance. Task demands are determined by the goals that have to be reached by task performance and cannot be linked directly to workload, which is subjective. Region merely indicates the interaction between performance

and workload. The same task can result in performance in region A2 for one individual, and may require effort compensation and thus region A3 performance for another. Also, in figure 2 the two types of effort compensation (Mulder, 1986, Cnossen, 1994) are split over two regions. In the A1 region deactivation is counteracted by state-related effort, while in region A3 task-related effort is exerted.

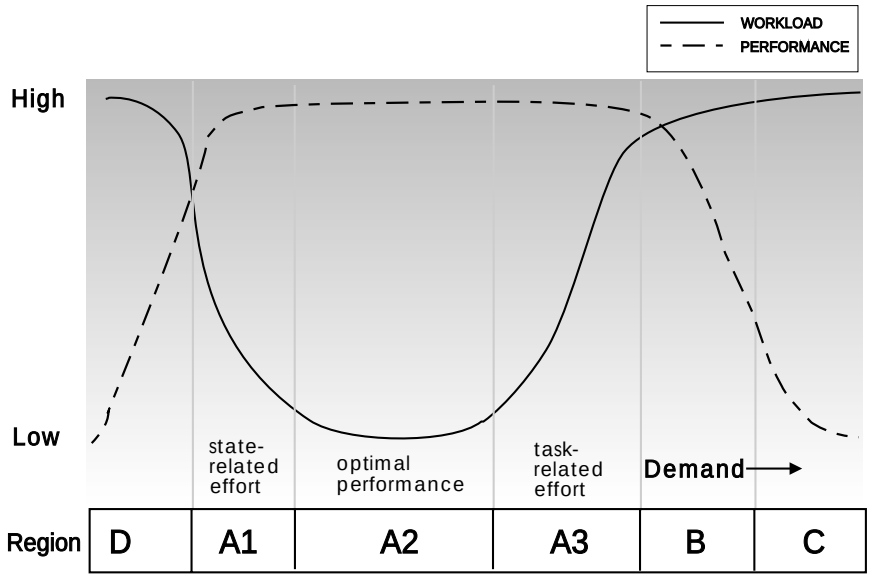


Figure 2. Workload and performance in 6 regions. In region D (D for deactivation) the operator's state is affected. In region A2 performance is optimal, the operator can easily cope with the task requirements and reach a (self-set) adequate level of performance. In the regions A1 and A3 performance remains unaffected but the operator has to exert effort to preserve an undisturbed performance level. In region B this is no longer possible and performance declines, while in region C performance is at a minimum level: the operator is overloaded.

In the model (figure 2) only one dimension of mental workload is displayed. What is depicted denotes the overall or sum relation between demand, workload and performance. The relation exists in principle for each separate resource. The implication is that auditory task demands, visual task demands and central demands do not necessarily have to be in the same region, which is in accord with Wickens' multiple-resource theory (Wickens, 1984).

With respect to the model the following questions can be asked: 'Which measure is sensitive when'? 'In order to assess mental workload, is one measure sufficient?' 'Do measures dissociate?' 'Can we deduce whether state-related effort or task-related effort was exerted, and if we can, how?'

This thesis focuses on how to measure driver workload. Different techniques, their characteristics and their use in applied settings, in particular in traffic research, will be evaluated. The technique's sensitivity in traffic research will be evaluated in the model on the basis of studies that my colleagues and I have performed. Particular attention will be paid to possible differential sensitivity of measures to increases in mental workload by changes in driver state opposed to changes in task complexity. The focus will be on performance measures that are specific for traffic research, a physiological measure (heart rate and its variability) and on two self-report scales. An overview of previously found results with respect to mental workload studies will be attended to before that, in chapter 3. But first the general properties of the measures will be considered.

3
Characteristics of measures

The measures that can be used for the assessment of mental workload have different properties. The properties range from very general aspects to very specific. A general aspect is, for instance, the amount of equipment that is needed. A more specific, and from a scientific perspective more important property is the validity of a measure. Is the measure reflecting the concept of mental workload as intended, or is it reflecting other concepts, e.g., physical workload? O'Donnell & Eggemeier (1986) categorize the criteria for the selection of a workload-assessment technique on the basis of the following properties of the technique: sensitivity, diagnosticity, primary-task intrusion, implementation requirements and operator acceptance.

Sensitivity

Is the technique able to reflect changes in workload? Sensitivity of a measure should be defined within region of performance. In the previous chapter a model that described the relation between workload, performance and demand was presented. A primary-task performance measure cannot possibly be sensitive to mental workload in region C or A, simply because in the region's definition included no change in performance. However, in the D and B regions changes in performance do reflect changes in workload. It is also likely that an operator is quite capable of indicating overload when demands are in the C region, and therefore a self-report measure's sensitivity can easily be different from performance measures per region. Evaluation of measures should therefore always be linked to the region of performance.

Diagnosticity

How capable is the measure in reflecting demands on specific resources? Diagnosticity is the ability to discern the type or cause of workload, or the ability to attribute it to an aspect or aspects of the operator's task (Wierwille & Eggemeier, 1993). A measure is said to be diagnostic within the context of the multiple-resource theory (Wickens, 1984) if it is sensitive to specific resource demands and not to others. Measures can be highly diagnostic and reflect a variation at a certain stage or on a certain locus of demand or they can be low on diagnosticity and reflect general demands. Pupil diameter is an example of a measure that reflects general demands and is low in diagnosticity. Pupil diameter is equally responsive to manipulations of different stages, such as response load or encoding and central processing load (Beatty, 1982). It is *not* sensitive to a specific type of resource expenditure. Other measures, e.g. some of the secondary task measures, are highly diagnostic. An example of a highly diagnostic measure is the evoked brain potential. The amplitude of the so-called P_{300}-component

of the evoked brain potential is sensitive to perceptual/central demands of a primary task (Gopher & Donchin, 1986). The choice for a diagnostic measure depends upon the measurement objective. If a general workload level has to be established, diagnosticity is not the most important selection criterium. If, however, the source of workload has to be traced a diagnostic measure can prove to be very useful and may guide to solutions of high workload demand.

Primary-task intrusion

The degree to which a technique degrades ordinary or primary-task performance is called primary-task intrusion. The disruption in ongoing task performance as a result of the application of the measurement technique is an undesirable property and should be minimized. Secondary-task techniques probably have the largest degrading effect on the primary task (Eggemeier et al., 1991). In particular the addition of an *artificial* secondary task may contaminate performance on the primary task. Self-report measures taken after completion of the task and most physiological measures seem to degrade primary-task performance the least.

Implementation requirements

Implementation requirements refer to practical constraints, such as the requirement of specific equipment or operator training. In field studies in particular, implementation requirements can become important. For example, the amount of equipment that is needed to measure eye movements may limit its use to laboratory settings (e.g. Unema, 1995). The same applies to the conditions in which some of the low-amplitude signal physiological measures can be properly assessed. Too much equipment might even result in primary-task intrusion.

Sometimes, in order to reach a stable or a reasonable performance level, subjects have to be trained extensively. In particular the requirement to obtain a reasonable dual-task performance can necessitate training. This is not necessarily a problem, but it does affect the time required before measures can be taken.

Operator acceptance

The degree of approval of the technique by the operator is referred to as operator acceptance. The operator's opinion about a measurement technique, especially the use of self-reports, largely affects the correctness and accuracy of the measure. In general acceptance is higher if the technique is less intrusive or artificial, while the face validity of specific measurements may enhance operator acceptance. If the use and usefulness of some of the measures is not clear to the operator, explanation about the measures' use is worthwhile and can help the operator to accept them (O'Donnell & Eggemeier, 1986). If the secondary-task technique is employed, operator acceptance is of primary importance. Acceptance can be enhanced by trying to let

the secondary task resemble activities that occur in the normal course of the operator's performance. In pilot performance, for example, activities like radio communication could be used (O'Donnell & Eggemeier, 1986)

Sensitivity, diagnosticity and primary-task intrusion are of major importance, while the latter two criteria, implementation requirements and acceptance, should be considered additional selection criteria. Some authors propose a slightly different categorization of criteria. Wickens (1992) added 'Selectivity' and 'Bandwidth & Reliability' to the list:

Selectivity

Selectivity is the selective sensitivity to mental workload and not to changes in such factors as physical load. Selectivity denotes the validity of the measure for workload assessment. A measure can be sensitive to mental workload only or be sensitive to other factors as well, in particular to physical load. If the measure is also sensitive to other factors, this may or may not be a reason to discard it for mental load measurement purposes, depending upon task and test environment. For instance, a measure that is sensitive to both physical and mental workload can be used as mental workload indicator when no physical effort is required.

Bandwidth and reliability

Bandwidth and reliability refer to the workload's estimate that has to be reliable both within and across tests. Stability of a measure between tests is what Wierwille & Eggemeier (1993) call 'transferability'. Measures that were developed in a laboratory setting do not have to indicate workload equally well in the field. Between applications much will depend upon the region of task performance. A measure sensitive to low levels of workload only will not be able to discriminate between levels within high demand situations. Comparison of results obtained in the different environment with samples taken from the same population should give a good estimate of reliability.

Interdependence characteristics

The above described characteristics are not independent of each other. A highly diagnostic measure is only sensitive to variations in workload in specific computational processes. Therefore diagnosticity restricts sensitivity. Also, diagnosticity presupposes selectivity. An other interrelation exists between bandwidth and sensitivity. Bandwidth is no more than the definition of the restricted area of test environments in which a measure is sensitive. Interaction between characteristics can be expected to be particularly high in the case of secondary tasks. For example, a diagnostic measure that is sensitive to secondary-task

performance can only be reliable if primary-task intrusion of the secondary task is low.

The most desirable characteristics of measures of mental workload are high sensitivity, preferably in a wide bandwidth, high reliability and low primary task intrusion. Diagnosticity can also be of major importance, in particular, if a certain stage of information processing is suspected to be affected.

The different measures and their characteristics will be discussed individually in the next chapter. Three groups of measures can be distinguished: self-reports of mental workload, task performance parameters and physiological indices. Overall experience with the measure's characteristics in the laboratory and field experiments will first be discussed, while in chapter 5 the range is narrowed to the applied domain of traffic research.

4
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

TLX, SWAT and MCH

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 (Eggemeier & 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 & Eggemeier, 1986, Eggemeier & Wilson, 1991 and Wickens, 1992, for overviews). Eggemeier & 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[3].

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

[3] 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 IBI[4], 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 (Porges & 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 2. Alternative naming of heart rate measures.

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

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 Hyönä 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 /EOG

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 & Seljos (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 *in*creases with task complexity. If the P_{300} is secondary-task-elicited it *de*creases 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_{300}s that are elicited by the irrelevant tones vary with primary-task workload in the same way as traditional secondary-task P_{300}s. 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, EMG

Research related to processing demands and mental effort and 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[5], 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

groups during research (cf. Wilson & Eggemeier, 1991, Sirevaag et al., 1993).

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 criterium 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.

5
Traffic research

Workload assessment is of interest to many applied settings ranging from VDT (visual display terminal) data-typing to space travel. Each area has its own specific problems. Much research was and still is performed in the area of aviation, in which, in particular, military flight studies have been carried out. In these studies pilots have to perform very complex tasks under extreme conditions and the selection of pilots is so stringent that in general only healthy young men and women are capable of realising these tasks. This selected group of people also serves as subject in aviation workload studies, where extreme G-forces and heart rate values of up to 160 bpm are no exception (see, e.g., Roscoe, 1993). These types of environments have a clear influence on physiological measures (e.g., on HRV quality, see Jorna, 1993).

No such forces are encountered in ground travel, though it could be argued that the human perceptual information system is unfit for the high speeds of travel possible in modern cars on motorways. Actual practice shows that this is not true and that many people are able to perform this task daily without negative consequences. Sometimes, however, our information processing system reaches its limit and things go wrong. In most of those cases a human error has occurred (Smiley & Brookhuis, 1987), resulting in a traffic-law violation that led to an accident (Rothengatter, 1991). In that case driving speed may have turned out to be too high to deal with safely. The selection criteria for driving are also far less strict than those applied in (military) flight, and, as a result, the population behind the wheel is far more diverse in capabilities. These factors, among others, make workload research in traffic an area with its own specific problems. Results booked in this area of research may benefit a large group of people.

Most workload measures used in traffic research have been developed, and tested, in the laboratory and in other applied settings such as the workplace (see e.g., Meijman & O'Hanlon, 1984) and aviation. The exceptions to these are the primary-task measures, for driving; the vehicle parameters.

A useful experimental design in traffic research is to compare task performance in an experimental (e.g., mental load) condition with performance under baseline conditions. A difference in performance can then be attributed to the experimental manipulation. Recently, however, Brookhuis (1995a, 1995b) has proposed critical levels of performance for different primary-task measures. These critical values can be considered performance margins as discussed in chapter 4. The criteria are not workload redlines, since they indicate the point at which performance should be considered to be affected, and thus indicate a shift from the A to the B region. Most of the measures' critical levels

have been linked to unsafe behaviour, e.g., a level at which the likelihood that the vehicle leaves the traffic lane increases to a major extent (see Brookhuis 1995ab). In the following evaluations the absolute criteria will be included.

Evaluation of workload measures on their characteristics in traffic research will mainly be restricted to work that my colleagues and myself have performed. Self-report measures, primary-task and secondary-task measures, and physiological parameters have been used in these field studies. From these studies, specific road sections or conditions were selected with increased task demands. The studies will be divided into two categories, studies that include an increase in complexity and studies in which driver state is affected. The first category can be further divided into two sub-categories, an increase in road complexity versus an increase in task complexity, i.e. the addition of a secondary task. Differential sensitivity of a selection of measures to mental load in relation to demand are of primary interest in the evaluations.

Selected sections or conditions

From one simulator and six field studies, experimental and baseline conditions or road segments (sections) were selected and the sensitivity of workload measures were compared between conditions and load categories. Sections were selected based upon expected effect of stressors or environment on workload, i.e. a selection based upon task demand. The following baseline and load conditions were selected:

Complexity studies - environment:

(1) From the '**weaving section study**' (appendix 1), a study performed on the A28-motorway, driving over the combined entrance/exit road-section was selected as experimental condition and compared with a baseline control section. In appendix 1 the load condition is referred to as 'ACC 2' (section ZL in the Dutch report). The baseline control section was a road segment with no entrance or exit and is indicated in the appendix as section 'CTR1' (C1 in the Dutch report). All subjects drove these sections two times, once without eye-movement registration equipment, once with the equipment mounted on their heads (indicated as 'c' for CEMRE, Continuous Eye Movement Registration Equipment).

(2) In the '**noise barrier study**' the same eye-movement equipment was used in one condition. Driving over a road section near a noise barrier was used as experimental condition and compared with driving along the same road section in the opposite direction, where no such barrier was present. The mid-part of the noise barrier was far closer to the motorway than the begin and end part. Driving on the motorway along the barrier created the impression that the screen 'approached' the car.

(3) In the '**road layout study**' (appendix 2) the effect of a changed road design of an 'A'-class road on driving behaviour, in particular on speed choice, and on mental load was tested. The baseline consisted of driving on an ordinary A-road section that either preceded or followed an experimental section. All roads had a speed limit of 80 km/h and were single carriageways with two lanes, separated by a white line. The experiment included roads in two environments: a road leading through a forest (Wr, woodland road) and a road leading through open moorland (Mr, moorland road).

Complexity studies - additional task:

(4) The '**car-phone study**' (appendix 3) was carried out both on a quiet motorway and on a busy four-lane ringroad. In the experimental conditions the drivers had to perform a difficult memory task, the PASAT, Paced Serial Addition Task (Gronwall & Sampson, 1974), while operating either a hand-held or a hands-free telephone set. In the experiment subjects drove and handled the car-phone five days a week for a total of three weeks. For the present comparisons, only data collected during the first week in which driver workload is likely to have been highest, were used.

(5) The '**tutoring**' or DETER (Detection, Enforcement and Tutoring for Error Reduction) study (appendix 4) is the only simulator study included in the comparisons. In this study, drivers had to complete four trials in a driving simulator where they drove through built-up areas, on A roads and on dual carriageways. The middle two trials, where an enforcement and tutoring system provided the subjects with feedback about detected violations, were compared with the first and last trial, when no feedback about violations was given. The tutoring messages and the required behavioural adaptation were suspected of increasing mental load.

If baseline performance in the above-mentioned studies is assumed to be in region A2, performance in the load condition with increased demand can be expected to be mainly situated in the A3 region (see figure 2), and perhaps in the neighbouring left-hand section of the B-region (Table 3). In some of the conditions, in particular the conditions that included the use of the CEMRE, mental load may have been additionally increased. The CEMRE reduced the visual field and subjects were therefore required to make additional head movements (see also appendix 1). However, in none of the studies is demand expected to be excessive.

Two studies in which driver state was affected were added. In the three load-conditions of the two studies, task difficulty was increased because driver state that was non-optimal as a result of the use of alcohol, a sedative drug or fatigue that followed lengthy driving. Performance in the load condition of these studies is expected to be, on average, situated in the A1 or D region (figure 2). The actual,

individual region of performance, however, will depend upon individual capacity, experience and goals set for performance.

Driver state studies:

(6) In the '**DREAM**' (Driver Related Evaluation And Monitoring) study (appendix 5) the effects of legally-allowed levels of Blood Alcohol Concentration (BAC ≤ 0.5 ‰), and the effects of fatigue (2.5 hours of driving, indicated in appendix 5 as 'vigilance'), were compared with baseline performance (the first hour of the last-mentioned condition). Driving on a busy four-lane ringroad and on a monotonous motorway were included in the study.

(7) Finally, in the '**antihistamine study**' the effects of a new-generation antihistamine (Ebastine) were compared with placebo and an active-drug control, Triprolidine. The active drug, which has a sedative effect, was chosen as the experimental condition and its effects were compared with the effects of placebo. In both conditions subjects had to drive on a busy four-lane ringroad and on a four-lane motorway.

Table 3. Traffic studies that are referred to in the figures in the following sections: region is the a priori and thus expected region of task performance as shown in figure 2. 'condition indicated' designates how the condition is referred to in the figures' legends, while the number of subjects is indicated under 'N'. References with a # are listed in full as appendix to this thesis.

study	test environment	selected load condition(s)	condition indicated	region	N	references
complexity, environment:						
Weaving section	On-the-road	combined entrance/exit	Weav	A3	52	De Waard (1991)#
		entrance/exit + Eye mark.	Weav(c)	A3-B		De Waard et al. (1990)
Noise barrier	On-the-road	Noise barrier	NoiseB	A3	22	Jessurun et al. (1990)
		Noise barrier + Eye mark.	NoiseB(c)	A3-B		
Road layout	On-the-road	Woodland Road, exp.	Wr	A3	28	De Waard et al. (1995)#
		Moorland Road, exp.	Mr	A3		Jessurun et al. (1993)
complexity, task:						
Car Phone	On-the-road	phone, motorway	Pmw	A3	12	Brookhuis et al. (1991)#
		phone, ringroad	Prr	A3-B		Brookhuis et al. (1989)
Tutoring	Simulator	warning messages	Tut	A3	27	De Waard et al. (submitted)#
						De Waard et al. (1994)
state:						
DREAM	On-the-road	Alcohol, motorway	Alc(mw)	D-A1	20	De Waard & Brookhuis (1991a)#
		Alcohol, ringroad	Alc(rr)	D-A1		De Waard & Brookhuis (1991b)
		Fatigue, motorway	Fat(mw)	D-A1		Brookhuis & De Waard (1993)
		Fatigue, ringroad	Fat(rr)	D-A1		Thomas et al. (1989)
Anti-histamine	On-the-road	Triprolidine, motorway	Tri (mw)	D-A1	15	Brookhuis et al. (1993)
		Triprolidine, ringroad	Tri (rr)	D-A1		De Vries et al. (1989)

In table 3 the above-mentioned studies are listed. In the following sections and figures the different selected conditions will be referred to as indicated in the column ‘condition indicated’. ‘N’ denotes the number of subjects that completed the tests.

Table 4. Measures used in each study (■). Measures will be explained in the next chapter. Alcohol and fatigue were conditions in one study, the DREAM study. SECOND. = secondary.

GROUP:	SELF-REPORT			PRIMARY TASK PERFORMANCE			SECOND. TASK PERFORMANCE			PHYSIOLOGICAL				
Measure:	RSME	RECL-A	ACTIV	SDLP	SDSTW	TLC	DELAY	MIRROR	EYEMOV	HR	HRV	.10 Hz	EMG	EEG
Study														
Weaving S.	□	■	□	■	■	□	□	■	■	■	■	■	□	□
Noise Barr.	□	■	□	■	■	□	□	□	■	■	■	■	□	□
Rd Layout	□	■	□	■	■	■	□	□	□	■	■	■	■	□
CarPhone	■	□	□	■	■	□	■	■	□	■	■	■	□	□
Tutoring	■	□	□	■	■	□	□	□	□	■	■	■	□	□
Alcohol	■	□	■	■	■	■	■	□	□	■	■	■	□	■
Fatigue	■	□	■	■	■	■	■	□	□	■	■	■	□	■
AntiHistam.	■	□	■	■	■	□	■	■	□	■	■	■	□	□

Which workload measurement method was used in which study can be seen in table 4. Three self-report scales were used of which two were unidimensional (RSME and Activation scale). The third scale, the activation scale of the RECL (Road Environment Construct List, see below), is based on multiple Likert-scales. As primary-task performance measures, the *SD* of the lateral position (SDLP), the *SD* of the steering-wheel movements (SDSTW), and the Time-to-Line Crossing (TLC)-measure were used. Mirror checking and delay in speed adaptations to a lead car’s speed changes in a car-following task are listed under secondary-task performance as embedded tasks. A genuine secondary task, performance on the PASAT, was only applied in the car-phone study. Three heart-rate measures are listed under physiology, average heart rate (HR), the modulation index of heart rate variability in the time domain (HRV) and variability in the frequency domain, in the 0.10 Hz band (.10 Hz). Activity of the facial corrugator muscle was used in one study, while ongoing EEG activity was used as physiological measure in the alcohol and fatigue (vigilance) conditions of the DREAM experiment.

The evaluation of measure sensitivity to workload, and in particular to differences in sensitivity to increased load in terms of

affected state opposed to increased complexity, will focus on the measures that were available in most studies, i.e.,

as *self-report measures*:
- RSME (Effort rating scale)
- Activation Scale

as *primary-task performance measures*:
- SD of the lateral position (SDLP)
- SD of the steering wheel movements (SDSTW)

as *physiological measures*:
- Average Heart Rate (HR)
- Heart Rate Variability in the time domain (HRV)
- Heart Rate Variability in the 0.10 Hz frequency domain (0.10 Hz)

For the sake of completeness not only the studies mentioned in table 3 will be evaluated, but other studies that were carried out in traffic and were found in the literature will, as far as possible and relevant, also be included in the next chapters.

5.1 Self-report measures

In this chapter, experience with driver self-report workload ratings will be described. The Dutch RSME and the originally German Activation scale will first be treated. The activation scale of the RECL is 'an odd one out', but is included because it was the only self-report rating that is available from the Weaving section, Noise Barrier and Road Layout studies (see table 4). Results obtained by others with the Task Load Index and SWAT are discussed under other self-report measures.

RSME, Rating Scale Mental Effort

In traffic research, the RSME (Zijlstra & Van Doorn, 1985, Zijlstra & Meijman, 1989) was used in the car-phone study and in the simulator experiment, and effects are compared with effects of the sedative antihistamine Triprolidine and the effects of alcohol and time-on-task (Car-Phone, Tutoring, Antihistamine and DREAM respectively in table 3). In figure 3 the absolute scores on the RSME scale of these four studies are indicated. Baseline ratings of effort of driving are compared with ratings of effort while driving and using a car-phone (load), driving without (baseline) vs. with (load) a switched-on enforcement and feedback system (Tutoring), and driving under placebo vs. under the influence of Triprolidine (load). The effects of 0.5 ‰ alcohol and fatigue (2.5 hours of driving) could not, due to the experimental design, be compared with baseline ratings, which could

not be collected. In figure 4 the change in scale values of the load condition opposed to baseline is indicated for the studies that included such a condition. All ratings were collected after completion of the driving task.

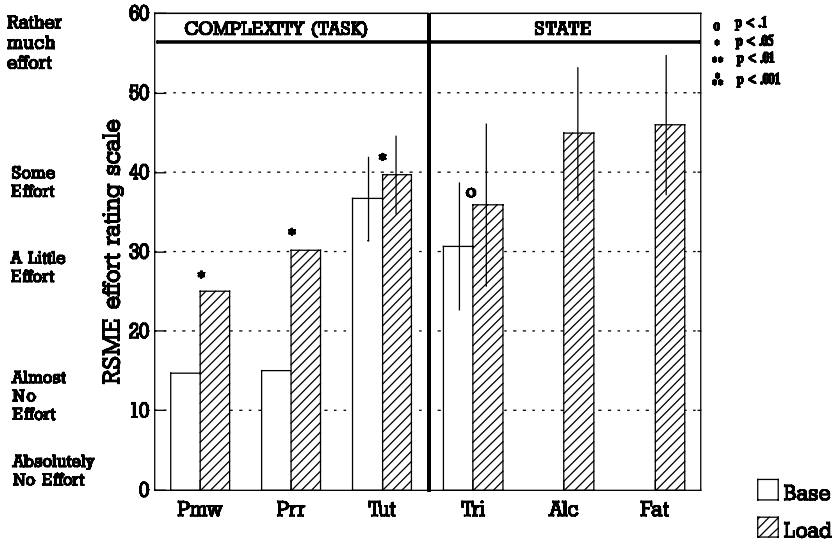


Figure 3. Average ratings of exerted effort on the unidimensional RSME of baseline driving and car-phone use (both on the motorway, Pmw, and on a busy ringroad, Prr), driving with and without an enforcement & tutoring system (Tut), driving under placebo and Triprolidine (overall rating, Tri), and driving on the motorway under the influence of alcohol (Alc(mw)) and while fatigued (Fat(mw)). If available, the 95% confidence interval is indicated.

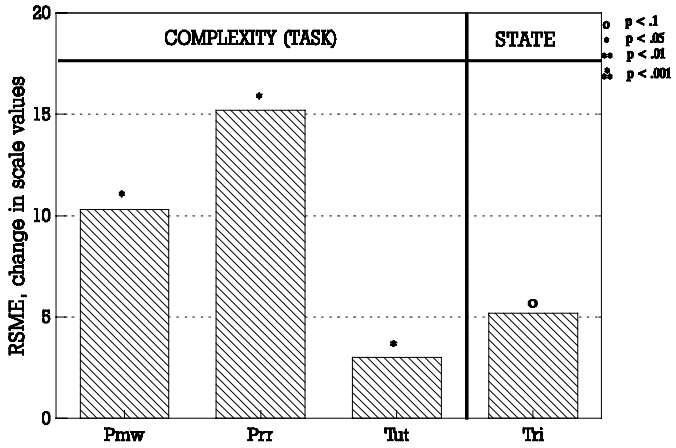


Figure 4. Average change in ratings of exerted effort on the unidimensional RSME in the case of car-phone use (Pmw and Prr), driving with an enforcement and tutoring system (Tut), driving under influence of Triprolidine (Tri), all compared with the baseline (control, placebo) measurement.

In all cases the RSME was able to distinguish between task-load situation and baseline. An increase in effort was reported in the case of car-phone use and as a result of the behavioural adaptation required by the enforcement system. The sedative effect of Triprolidine also resulted in an increase in effort exerted. Between the Tutoring and Car-phone study important differences in baseline values were found. These differences may reflect differences between the subjects who participated, but it is more likely that they reflect differences between the baseline tasks. As mentioned previously, the effects of task load were compared with baseline driving. For the Tutoring experiment, baseline driving included handling a simulator car and driving through a varied area, while in the Car-phone experiment an instrumented vehicle had to be driven through traffic. Judging from the absolute scores, the latter task is less effortful. Recently, support for this statement was found in a study in which the same subjects performed the same task both in traffic and in a simulator (De Waard & Brookhuis, in press). Driving in the simulator required more effort as measured with the RSME.

Activation scale

Bartenwerfer's activation scale was used in two studies that are listed in table 4. In the DREAM experiment, however, no baseline ratings could be collected[6]. The effect of Triprolidine on reported activation level estimated over the whole journey, was not significant. The application of the activation scale to traffic research has mainly been limited to drug research. An indication of the measure's sensitivity to affected driver state can be obtained by looking at the results from these 'drugs & driving' studies. In figure 5 the change in scale values, compared with placebo, is listed for drugs as measured in five on-the-road studies. The average placebo value over all studies was 131, which is just below the reference point 'I am solving a crossword puzzle' (see appendix B for the scale). Data regarding antidepressants, hypnotics, analgesics, tranquillizers and antihistamines have been taken from Louwerens et al. (1983), Volkerts et al. (1984), Brookhuis et al. (1985a), Volkerts et al. (1987) and De Vries et al. (1989), respectively.

The most pronounced effect on reported activation level was the reduction found in the antidepressant study. One hypnotic reduced reported activation level, while the analgesic showed a dose-related effect. This last effect was not in the expected direction, activation level increased with an increase in dose of this drug for pain-treatment. However, in that study performance measures did not decline with increasing dose either, and nor did reaction-time performance in a laboratory task (see Brookhuis et al., 1985a).

[6] Average rating for Alc(mw) was 130.5 and for Fat(mw) 139.0.

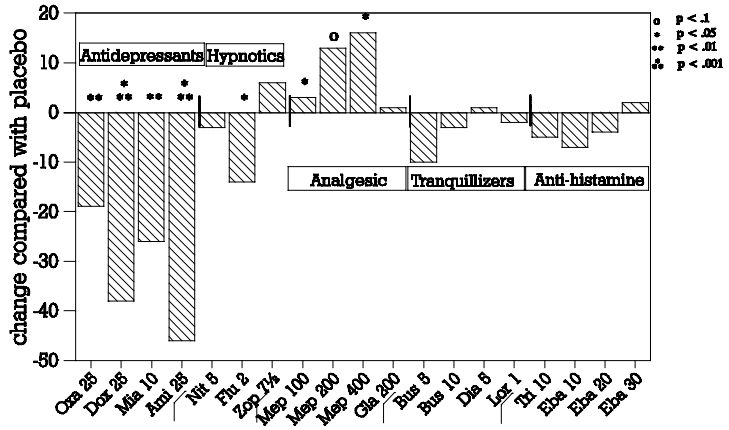


Figure 5. Average change in rated activation in five drug studies compared with the placebo conditions

No effects on the scale were found in the tranquillizer and antihistamine studies. On the basis of these studies it seems that the scale is fit to be used for effects on subjectively experienced effects on the Central Nervous System. The relation of the scale to mental workload in general is, at present, hard to assess. It seems likely that the scale is of particular use in the areas further away from optimal performance, hence in the D and C regions.

Other self-report measures used

RECL, Road Environment Construct List

The Road Environment Construct List (RECL, Steyvers, 1993, Steyvers et al., 1994) was developed to measure appraisal of road environments. The RECL is a three factor scale. Each of the sixteen items load on one of three factors. The factors are: 'Hedonic value', which denotes the aesthetic appraisal of the road and its environment, 'Perceptual variation' denoting the heterogeneity in the road environment, and 'Activation value' denoting the extent to which the road and environment are considered to be activating. The latter factor may be useful for workload measurement in a traffic environment.

The RECL was used in studies in which the RSME was not used and therefore the Activation value of the RECL is included in the evaluation on usefulness as an indicator of driver activation. Though the driver is asked to evaluate the road and its environment, an activating effect of the environment could be related to road-environment demands and might therefore influence driver mental activation.

Although the trend in scores of baseline and load conditions in the road-layout experiment was in the direction of increased load, differences between the two conditions on both roads (Wr and Mr) were not significant. In the two motorway studies no baseline measurements were taken. However, two other conditions of these experiments could be compared: driving without and with ('c') eye-movement equipment mounted on subjects' heads. In both studies subjects did not rate the activating influence of the environment different as a result of the equipment (see table 5).

Table 5. Average rating on the Activation scale of the RECL. Baseline measurements were only collected in the road layout study.

	Baseline	Load	Significance (t-test)
Weaving Section	-	3.6	ns
Weaving Section (c)	-	3.7	
Noise Barrier	-	3.6	ns
Noise Barrier (c)	-	3.7	
Road Layout Wr	4.2	4.5	ns
Road Layout Mr	3.2	3.8	ns

Other self-report measures in other studies

TLX, Task Load Index

In none of the studies listed in table 3 was the NASA Task Load Index (TLX) used. A few on-the-road studies reporting the use of this self-report measure were found. Fairclough et al. (1991) used the RTLX (Byers et al., 1989) in a dual-task performance study. They found an increase in overall workload in the dual task condition, which consisted of driving plus having a conversation, compared with single-task performance, which was normal driving. The RTLX was also used in another study performed in the same vehicle (Vaughan et al., 1994). In the experiment RDS (Radio Data System) messages had to be attended to. The messages were presented to subjects in three conditions in a within-subjects design: 1. auditory, 2. auditory and continuously visible on a display, and 3. auditory and temporarily (15 s) visible on a display. Overall RTLX mental workload rating was lowest for condition 2, auditory plus visual constant. The RTLX factors 'mental effort' and 'time pressure' showed a similar effect (the lowest rating for condition 2, the highest rating for condition 1 and slightly less high for condition 3). The results found in this dual task study illustrate the diagnosticity of the RTLX in the reflection of higher scores on the time-pressure factor in the case of auditory messages and no or quickly disappearing visual information.

In a simulator study in which the effects of a hands-free car-phone were tested Alm & Nilsson (1994) found an effect of the car-phone task on all subscales of the TLX. An interaction between car-phone use and driving-task difficulty (in terms of driving a straight

opposed to a winding road) was only found on the frustration subscale, and not on the mental-demand or operator-effort subscales.

SWAT

The SWAT was used in simulator and on-the-road experimental tests of the GIDS system (Janssen et al., 1994). The system gave support to the driver by route guidance messages, and with respect to speed, collision avoidance and lane keeping (simulator trials only). Judging from the SWAT-reference that was provided in the text, an adapted version was used in which the card-sort section was left out. The authors report the overall mental workload index, which is defined as the addition of three 3-point scales (time stress, mental effort and psychological stress) resulting in a sum-scale range from 3 to 9. SWAT ratings differed between integrated and non-integrated GIDS support both in the simulator trials and in the on-the-road tests. The difference between integrated and non-integrated support was that support was only scheduled according to demand in the first condition. Scheduling includes, for instance, postponing an incoming phone call in the event that a lead vehicle brakes suddenly.

In an on-the-road experiment Verwey & Veltman (1995) found that summational SWAT ratings were equally sensitive to increases in workload as ratings on the RSME. Inclusion of the card-sort task for SWAT did not yield more accurate workload estimates.

Properties of self-report measures

Sensitivity, selectivity, diagnosticity, validity and primary-task intrusion are of major importance for a measure of driver workload. These properties were assessed as adequately as possible on the basis of the above-described experiments. The region in which the measure was found to be sensitive is indicated under sensitivity, and region-sensitivity has to be considered the prime property.

The RSME is designed to reflect operator effort. In the car-phone and tutoring experiments the RSME was found to be sensitive to task-related effort, while in the antihistamine study the rating scale was sensitive to state-related effort. Accordingly, when performance is in Region A1 and A3/B the RSME can be expected to reflect driver mental effort. The drug studies showed that the activation scale is in particular sensitive to an affected driver state as a result of (highly) sedative medicine such as hypnotics and antidepressants. Increased activation levels, e.g., as a result of the use of amphetamine (Sanders, 1983), can be expected to be reflected in higher activation scores, but as yet, there is, to my knowledge, no evidence available from empirical studies to support this prediction.

Diagnosticity for the two unidimensional scales is low unless they are applied per task dimension as proposed by Zijlstra & Meijman (1989). Selectivity is difficult to assess as the main other factor to which the scales could be sensitive, physical workload, is very restricted in driving. Reliability is high, as sensitivity to mental

workload in the different studies is high. Primary-task intrusion is low as long as the ratings are asked after completion of the task. Since hardly any equipment is required for collection of the measures the implementation requirements are low. No problems in operator acceptance have been encountered, so informal evidence supports high operator acceptance. In table 6 the results are summarized.

Table 6 Summary of properties of self-report workload measures.

	Measure	
Property	RSME	Activation
sensitivity (Region)	(D-)A1, A3-B	D, (B-C)
diagnosticity	low	low
selectivity	prob. high	(?)
Reliability	high	high (?)
primary-task intrusion	low	low
implementation requirements	low	low
operator acceptance	high	high

5.2 Primary-task performance measures

Parkes (1991) defined the primary task of the driver as maintenance of safe control over the vehicle. One of the major subtasks in vehicle control is lateral position control. Therefore, a measure of driving deviations from the centre of the lane is a good means to assess primary-task performance in car driving. Lateral deviation, or more specifically the *SD of the Lateral Position* (SDLP), has been shown to be a sensitive performance measure (e.g., Hicks & Wierwille, 1979, O'Hanlon et al., 1982, O'Hanlon, 1984, Brookhuis et al., 1985b, Green et al., 1993b). The task of keeping a vehicle between the lines of a lane is largely a psychomotor task involving eye-hand coordination. The term 'tracking-ability' is sometimes applied to it (e.g., Stein et al., 1987), stressing the strong resemblance to the laboratory task.

Standard Deviation of the Lateral Position

In figure 6 the average (right-hand lane) SDLP in baseline and load conditions is displayed, while in figure 7 the change in *SD* of the lateral position compared with baseline is shown. This relative measure was added to neutralize the differences in baselines between studies, which are likely to have been caused by differences between roads/road segments, season (weather) and so on. The absolute value of the SDLP in the Tutoring experiment is omitted from figure 6 because in the experiment the road width and test environment were very different from the on-the-road tests. In both figures the critical impairment levels

(Brookhuis, 1995a, 1995b) are indicated, while in figure 7 the impairment in lateral position control found in an 'alcohol calibration study' (Louwerens et al., 1987) is included.
An increase in SDLP, i.e. an increase in swerving, was found near the noise barrier (but only in the condition without eye-movement measurement), and as a result of alcohol (Alc(mw)) and prolonged driving (Fat(mw)). A decrease in the SDLP in the mental load condition was found in conditions in which subjects handled a car-phone (Prr and Pmw), when the enforcement system was switched on (Tut), and on the experimental road-layout (Wr and Mr). In some cases, the relative short section that was selected as load condition could have had an effect on SDLP. Near the noise barrier, for instance, the average lateral position on the road moved to the left. In the road-layout experiment the road surface and effective road width had been reduced, forcing drivers into more accurate lane-keeping. The effect of lane width on tracking performance was also found in a pilot study performed in a driving simulator (Green et al., 1993b), they found an increase in SDLP with increases in lane width. Taking these factors into account leaves only primary-task performance *decrements* on the SDLP measure as a result of alcohol and prolonged driving. In the Tutoring and Car-phone experiment primary-task performance under mental load, as measured by SDLP, even improved, while the sedative drug Triprolidine and driving on the Weaving Section did not lead to a significant increase in SDLP.

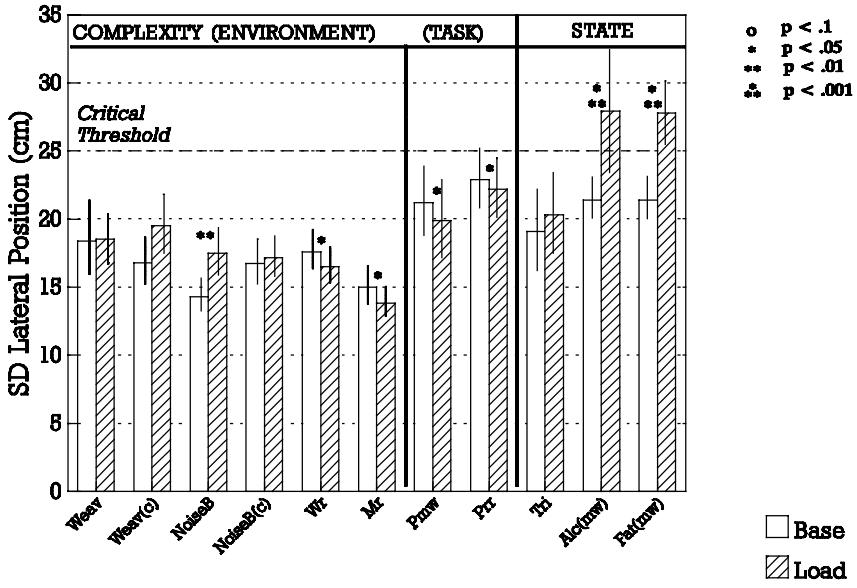


Figure 6. Standard deviation of the lateral position under baseline and mental load conditions. The studies from which the conditions were selected are listed in table 3. The indicated absolute threshold indicates driver impairment (see Brookhuis, 1995ab). The 95% confidence interval is also indicated.

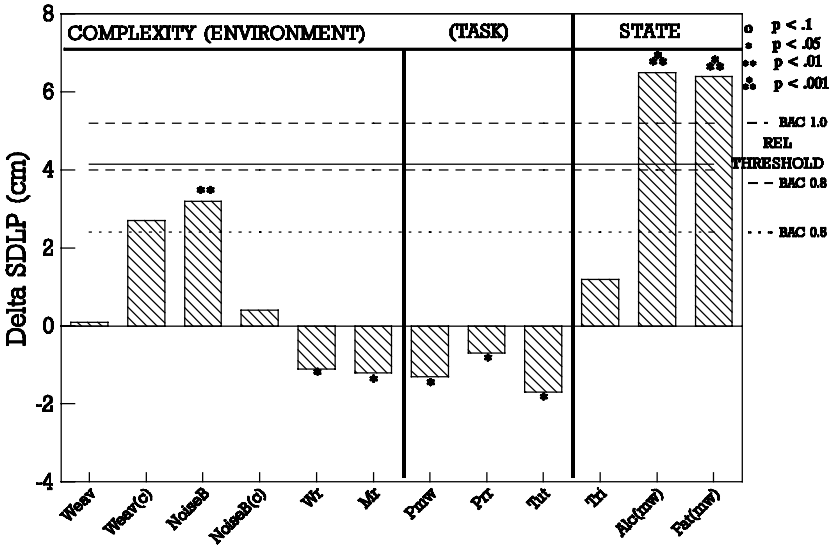


Figure 7. Change in standard deviation of the lateral position under mental load compared with baseline measurements. The studies from which the conditions were selected are listed in table 3. The indicated relative threshold denotes driver impairment (see Brookhuis, 1995ab). The indicated BAC (Blood Alcohol Concentration) values are impairment levels as found in an 'alcohol calibration study' (see Louwerens et al., 1987)

Standard Deviation of the Steering wheel movements

Related to the SDLP, but closer to one of the main sources of swerving, is the driver's steering behaviour. Due to relatively low-attentional driving demands, or due to attentional demands of additional tasks, drivers do not pay continuous attention to the lane-tracking (steering) task. This results in steering 'holds', i.e. periods without steering-wheel movements (see Macdonald & Hoffmann, 1980, Godthelp et al., 1984). Several steering measures have been developed, from relatively simple measures, such as the number of zero-degree crossings of the steering-wheel or steering-reversal rate (McLean & Hoffmann, 1975), to more complex measures involving frequency analyses (McLean & Hoffmann, 1971, Blaauw, 1984) and compound functions (Fairclough, 1994). Steering-reversal rate (McLean & Hoffmann, 1975, Macdonald & Hoffmann, 1980) and the *SD* of the steering-wheel movements, always measured on straight road segments, are frequently used performance measures that are not complicated to calculate. In the figures 8 and 9 the (Δ) SD of the steering-wheel movements (SDSTW), on sections with hardly any or no curvature is shown. Again the critical impairment level (Brookhuis, 1995ab) is displayed in both figures. In three studies the SDSTW increased in the load condition, in two studies to a level above the absolute impairment criterium. The elevated SDSTW at the experimental road-layout (Mr)

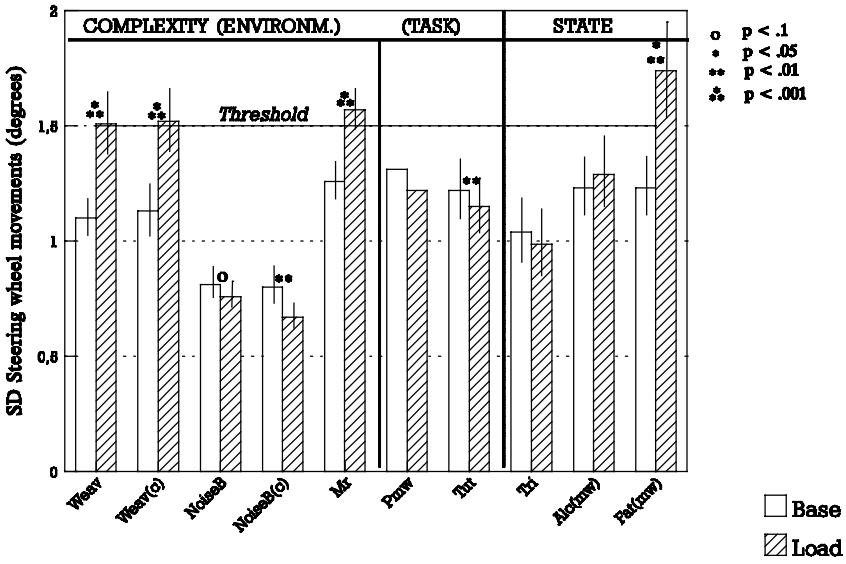


Figure 8. Standard deviation of the steering-wheel movements under baseline and mental load conditions. The studies from which the conditions were selected are listed in table 3. The critical threshold level indicates driver impairment (see Brookhuis, 1995ab). If available the 95% confidence interval is displayed.

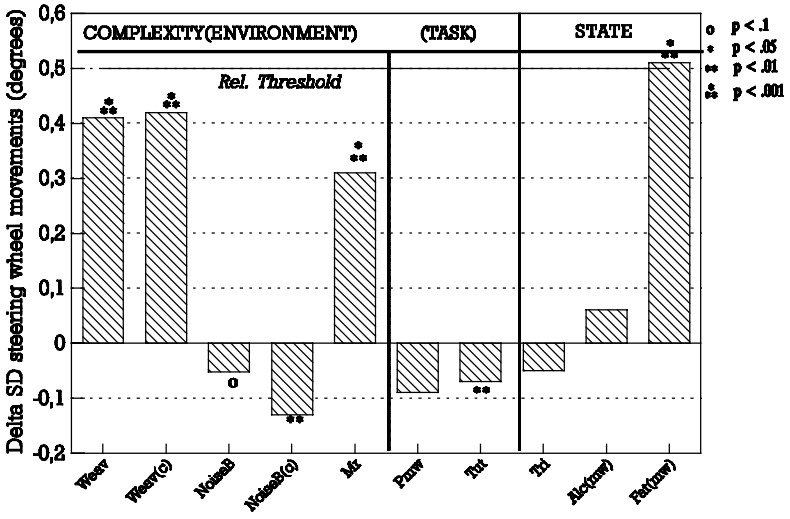


Figure 9. Change in standard deviation of the steering wheel movements under mental load compared with baseline measurements. The studies from which the conditions were selected from are listed in table 3. The indicated relative threshold indicates driver impairment (see Brookhuis, 1995ab).

was unexpected. However, the two selected road segments may have differed slightly in curvature. The experimental road section was somewhat more curved. Road curvature may have had a similar effect on the SDSTW in the experiment that focused on the Weaving Section. In the other experiments completely straight or even the same motorway sections were compared with each other, which in general is to be preferred. In the simulator (Tut) and Noise Barrier study a significant decrease in SDSTW was found. A decrease in SDSTW may be indicative of increased steering effort, and thereby of more accurate steering, e.g., as a result of road environmental demands.

A combined statistical test

The statistical power of the individual tests can be increased by combined testing of the effects found in the different experiments. If it is assumed that it is the same parameter that is affected in the different studies (and that that parameter is mental workload), then the effects found in the studies can be tested in combination by (Snijders, 1995):

$$z = \frac{\sum_{i=1}^{k} \alpha_i \theta_i}{\sqrt{\left(\sum_{i=1}^{k} \alpha_i^2 \, SE_i^2 \right)}}$$

with k = the number of experiments
θ_i = the estimated effect in experiment i
SE_i = standard error in experiment i
α_i = weight of experiment i.

z is tested in a standard-normal distribution, with H_0: $\theta = 0$.

This test was applied to the SDLP and SDSTW measures. The following results were found:

SD of the lateral position:

Complexity[a,b]:	*z* = -0.29, NS
Complexity (environment)[a,c]:	*z* = +1.87, $p < 0.05$
Complexity (task)[b]:	*z* = -2.63, $p < 0.005$
State[c]:	*z* = +4.51, $p < 0.0005$

[a] = Without Wr and Mr due to reduced road width.
[b] = Weighted, $\alpha = 1$ for Prr and Pmw, $\alpha = 2$ for Tut.
[c] = All conditions equal weights.

The road layout experiment was excluded from the tests as changes in SDLP cannot be solely attributed to changes in mental workload, but are combined with effects of reduced road width. In the Complexity tests the two car-phone conditions were weighted to balance effects with the single condition of the tutoring experiment.

Increased complexity in terms of a change in environment as opposed to additional tasks have dissimilar effects on SDLP. Increased task complexity concur with reduced SDLP, while increased environmental demands coincide with an increase in SDLP. Tested together as effect of 'complexity' levels out effects and renders a nonsignificant result. These results will be further discussed in chapter 5.5.

SD of the steering wheel movements

Complexity[a]:	$z = +3.86$, $p < 0.0005$
Complexity (environment)[d]:	$z = +8.35$, $p < 0.0005$
Complexity (environment)[a]:	$z = +4.93$, $p < 0.0005$
Complexity (task)[e]:	-
State:	$z = +16.45$, $p < 0.0005$

[a] = Without Mr due to reduced road width in load condition.
[d] = Mr is included in this test with $\alpha_{Mr} = 2$, while $\alpha_{Weav} = \alpha_{Weav(c)} = = \alpha_{NoiseB} = \alpha_{NoiseB(c)} = 1$.
[e] = Not tested, only standard error information from one study (Tut) available

---

An increase in complexity of the environment and a decreased driver state both lead to a significant increase in the SD of the steering wheel movements. Increased task complexity reduces the SD of the steering wheel movements. These results will, together with the effects on SDLP, be discussed in chapter 5.5.

Other primary-task performance measures

Time-to-line crossing

While the SDLP and SDSTW mainly reflect performance at the control level, one level higher, at the manoeuvring level of performance, the *Time-to-Line Crossing* (TLC, Godthelp 1984) is a measure of driver primary-task performance. TLC is a continuous measure that represents the time required for the vehicle to reach either the centre or edge line of the driving lane if no further corrective steering-wheel movements are executed. TLC reflects the amount of time drivers can neglect path errors. Due to the measure's skewness, in general minimum, median or 15% TLC values are calculated (Godthelp et al., 1984, Godthelp, 1988). TLC is expected to reflect driving strategy and in particular occlusion strategy (time spent not looking at the road). With increases in mental-load, smaller TLC values can be

expected; a more demanding task is likely to decrease the amount of time spent looking at the road.

In table 7 median and minimum TLC, as well as the change in TLC relative to baseline are depicted for the DREAM (for TLC see De Waard & Brookhuis, 1991b) and road-layout study. In the vigilance condition a decrease in TLC was found. This is in accord with the increase in number of steering-wheel holds that was found as a result of time-on-task (De Waard & Brookhuis, 1991b). In the road-layout study the layout of the road had been changed significantly. Drivers were more or less forced to drive close to the centre line and as a result the left-hand TLC decreased, while the right-hand TLC increased. This measure actually reflects the time required to reach an *imaginary* edge line, as the line had been removed! As a result, interpretation of the TLC measures in terms of mental load measures is not useful with data of the road-layout study.

Table 7. Median and minimum time-to-line crossing (s) in baseline and mental load conditions. Change in TLC denotes the change from baseline to load. Significant results have been printed in **bold**.

	left hand Median TLC		left hand Minimum TLC		right hand Median TLC		right hand Minimum TLC	
condition	base	load	base	load	base	load	base	load
Mr	4.95	**4.27**	2.35	**1.91**	3.10	**3.59**	1.19	**1.50**
Alc(mw)	5.31	5.42	1.89	1.83	3.79	3.98	1.58	1.69
Fat(mw)	5.31	**3.95**	1.89	**1.71**	3.79	**3.11**	1.58	**1.30**

	Median TLC Change in TLC		Minimum TLC Change in TLC	
	left	right	left	right
Mr	**-0.68**	**0.49**	**-0.44**	**0.31**
Alc(mw)	0.11	0.19	-0.18	0.11
Fat(mw)	**-1.36**	**-0.68**	**-0.07**	**-0.28**

Results with respect to TLC, SDLP and SDSTW from other author's studies

Riedel (1991) also used the TLC measure in a drug study in which subjects performed a driving task on the road. He found a maximum *in*crease in median TLC (undifferentiated to line) of 0.15 s in the Triazolam condition, while baseline median TLC on the motorway was 4.69 s. SDLP in the same condition increased with 6.6 cm to 30.7 cm. On the basis of these data, he concluded that SDLP is the most sensitive measure for driver impairment.

The effect of Blood Alcohol Concentrations on SDLP as found by Louwerens et al. (1987) have been indicated in figure 6. The sensitivity of the measure to an affected driver state as a result of the

use of hypnotics are summarized in Brookhuis (1995b). Significant increases in SDLP starting at 2.5 up to 7 cm are reported.

Van Winsum et al. (1989) compared steering-wheel movements of drivers who navigated from a map with the steering-wheel movements of drivers who were guided vocally. They found no effect on steering-wheel movements in the more demanding map condition. This result may be related to the urban road environment. It is likely that the use of most primary-task control indices (SDLP, SDSTW) is confined to non-urban environments. In urban traffic most steering-wheel movements will be related to longitudinal and lateral tracking demands (Wildervanck et al., 1978).

Green et al. (1993a) compared driving behaviour and self-report ratings of difficulty of route guidance messages using three different interfaces. Only slight differences in *SD* of steering wheel movements were found, the largest *SD* of steering wheel movements were measured when the information was displayed in the instrument panel (1.1˚), followed by a simulated Head-up display (1.0˚). The *SD* of the steering wheel movements were smallest (0.9˚) for auditory presented information. Ratings of difficulty of use of the route guidance information while driving that were given after the test rides (Green et al, 1993a, p.82) followed the same pattern, the lowest difficulty rating being given for the auditory information. However, memory load in the case of auditory route guidance was largest. In all three conditions route guidance information was additionally combined with information regarding vehicle state and traffic information that was presented to the driver in the instrument panel at a different location. This additional information could have interacted with the route guidance messages and therefore a relation between type of interface and mental load is hard to assess accurately.

Fairclough (1994) measured steering-wheel movements in a study in which subjects drove under the influence of low amounts of alcohol, and under placebo conditions. Just as in the DREAM study (see figure 8) he found an increase in the standard deviation of steering-wheel movements of drivers with a BAC up to 0.5 ‰.

Other primary-task measures in other studies

Apart from the above-mentioned *accuracy* measures in vehicle control, sometimes *speed* measures are used in the assessment of primary-task performance. An example of a speed measure is the time that is required to finish a route. Both Jordan & Johnson (1993) and Fairclough et al. (1991) found the time required to complete a route to be significantly longer in the load condition in which subjects had to adjust a stereo or had a conversation, compared with normal driving along the same route. The measure can be indicative of a strategic choice for a lower driving speed to compensate for high information load, and accordingly lead to a decrease in mental load. Similar compensatory strategies are reported for slower decision making and

slower action performance in elderly drivers (Brouwer & Ponds, 1994). Brown et al. (1969) also found an increase in time required to finish driving a circuit as a result of the use of a car-phone, while Van Winsum et al. (1989) found the same effect -an overall lower driving speed- when they compared map navigation with vocal-route guidance. However, the application of the measure is rather rough, and in non-controlled environments, e.g. in on-the-road studies, the measure is susceptible to disturbance factors such as traffic density. The use of speed measures as a sensitive indicator of increased mental load seems, therefore, to be the most reliable in laboratory and simulator experiments.

Properties of primary-task performance measures

Lane-keeping in experienced drivers is to a large extent determined by automatic, control-level processing. Consequently, measures of accuracy in lane-keeping, such as the SDLP and SDSTW, would not be expected to be sensitive to variations in mental workload in the A-region. The different experiments, however, show that this is not the case, both SDLP and SDSTW being sensitive measures. A likely explanation for this is that there is no 'pure' automatic and controlled behaviour, but that aspects of automatic behaviour remain influenced by controlled processing (Schneider and Fisk, 1983). Strategy sets performance margins and the inaccuracies that are allowed. This also clarifies why improvement on these primary-task performance measures is possible. Increased task demands can lead to increased driver effort, which increase primary-task performance if under baseline conditions inaccuracies are allowed. This issue will be further discussed in chapter 5.5.

Although improvement in primary-task performance measures is possible, in general, affected task performance implies reduced task performance, and this is the case in the D, B and C regions. As task performance is at a minimum level in the C-region, performance measures will no longer vary with changes in demand in that region. Sensitivity of the SDLP and SDSTW is highest in the B and D regions. In studies in which driver state was reduced, a decrease in SDLP and SDSTW was found. The same is true for the increased environmental demand studies. Diagnosticity of the measures is low, although the difference in direction of the effect as found between Complexity environment vs. Complexity task may be an indication of differential sensitivity. Selectivity is hard to assess on the basis of the driving studies reported above. Hardly any physical effort is required in driving, and emotional stress, for instance, was not tested. It is quite possible that the measures are affected by these factors and therefore selectivity is expected to be relatively low. Sensitivity to mental workload as found in the different tests results in a 'high' rating for reliability. The implementation requirements for the measurement of steering wheel movements are low. A potentiometer mounted on the steering wheel

column with a measurement range of 90˚ (± 45˚) and a resolution of 0.1˚ is adequate for accurate measurement of movements on noncurved road sections. For the measurement of the vehicle's lateral position more complex equipment is required. A useful device is the so-called 'lane tracker', which resembles a camera but the interior consists of an array of diodes that are sensitive to differences in light intensity. The camera is directed towards the road delineation (see appendix 5). A relatively cheap but labour-intensive solution is to make video registrations of the road scene (De Waard & Steyvers, 1995). The advantage of the latter technique is that it can also be applied on roads without delineation. In the future progress in camera techniques will probably facilitate automatic detection of road delineation or road shoulder. Operator acceptance of the measures is high because registration is unobtrusive. Table 8 provides an overview of primary-task measures' properties.

Table 8. Summary of properties of primary-task workload measures.

	Measures	
property	SDLP	SDSTW
sensitivity (Region)	D, B	D, B
diagnosticity	low	low
selectivity	(low)	(low)
reliability	high	high
primary-task intrusion	none	none
implementation requirements	high	low
operator acceptance	high	high

5.3 Secondary-task performance measures

If no specific instructions are given it is not clear which task is given priority. In heavy traffic the conversation with a passenger will probably be disrupted to maintain driving performance while in quieter environment and during a very interesting conversation driving performance will be affected (Wickens, 1984). Moreover, while the division between primary and secondary tasks may be very clear-cut for most laboratory tasks, this is not the case in driving. In traffic, behaviour is quite often related to the manoeuvre that is performed. Monitoring of rear traffic can be crucial if an overtaking manoeuvre is planned. In those cases the task of looking into mirrors and over one's shoulder cannot be called 'secondary'. Task integration can also blur the transition from primary to secondary task. A good example of dual-task integration is *car-following*. In heavy traffic this task will be added to the primary task of lateral and longitudinal vehicle control. It is the addition of a task, but the added task is not artificial. The experience of

various subtasks as a ‘single task’ is in particular likely if the subtasks are related or coherent (see, e.g., Korteling, 1994ab). Viewed in this way, car-following performance could be an embedded secondary task. However, a condition for a task to be termed embedded is that it is given lower priority than the primary task. It is not certain that car-following is given lower priority than lane-keeping. Perhaps a useful description of a secondary task in traffic research is that the task does not have to be performed continuously. In this way, the primary task remains restricted to speed and lateral vehicle control. Secondary tasks are non-continuous tasks, i.e. headway keeping can only be performed in case a lead vehicle is present and looking into the mirrors is performed at intervals. The definition is weak, but so is the separation of primary and secondary tasks in traffic.

Car-following

At the Traffic Research Centre a car-following task for use in real traffic has been developed (Brookhuis et al., 1994). In the task, a lead car’s speed fluctuations have to be followed by the driver of an experimental vehicle. This task is designed to be sensitive to impairment of performance in attention and perception, while lane-tracking is merely sensitive to performance on eye-hand coordination. In terms of the hierarchical model of car driving (Janssen, 1979, Michon, 1985, see also chapter 1) the lane-tracking parameters (SDLP, SDSTW) reflect performance at the control level, while car-following parameters reflect performance at the manoeuvre level. The main parameter in car-following performance is the delay in reaction to speed changes of the lead vehicle. We (Brookhuis et al., 1994) obtain this measure by performing a coherence analysis on the speed signals of the lead and the following car. Apart from delay (calculated as ‘phase shift’ between the two speed signals in the frequency domain) two other parameters are computed, which both give an indication of ‘how well’ the car-following task is performed. Coherence is a measure of the accuracy of car-following performance, while the modulus indicates the amount of overreaction to speed changes by the following car (Porges et al., 1980).

The car-following task was included in the car-phone, DREAM and antihistamine studies. Delay increased in conditions in which a car-phone was used (+23%), after alcohol consumption (+19%), and in the condition in which Triprolidine had been taken (+42%). Time-on-task (Fatigue) did not affect delay, but coherence slightly decreased in this condition.

Mirror checking

Mirror checking is another good example of an embedded secondary task that is specific for car driving. Two variables can be distinguished in mirror checking: frequency and duration. Total *duration* of mirror checking was measured both in the Weaving Section and the Noise Barrier study. In the Noise Barrier study, however, only data related to the load condition were available. In this condition no more

than 2.7% of the total time was spent looking in the mirrors. In the Weaving Section study, the difference in mirror-looking time between load (10.6%) and control (10.2%) was not significantly different. In an in-vehicle navigation study, Fairclough et al. (1993) compared driving performance and visual attention while navigating from map vs. from a text-LCD screen. They found a decrease in duration of fixations in the rear-view mirror in the higher demand (i.e., map) condition. In another study, reported in the same paper, glance frequency (but not glance duration) in the rear-view mirror was decreased in the condition in which internal vehicle 'checking behaviour' of a display was higher. The authors' conclusion was that glance duration and glance frequency are representative for different aspects of driver behaviour. Duration appears to be sensitive to difficulty of information intake, while glance frequency represents visual activity in terms of checking behaviour, both inside (e.g., speedometer checking) and outside (e.g., mirror checking) the vehicle.

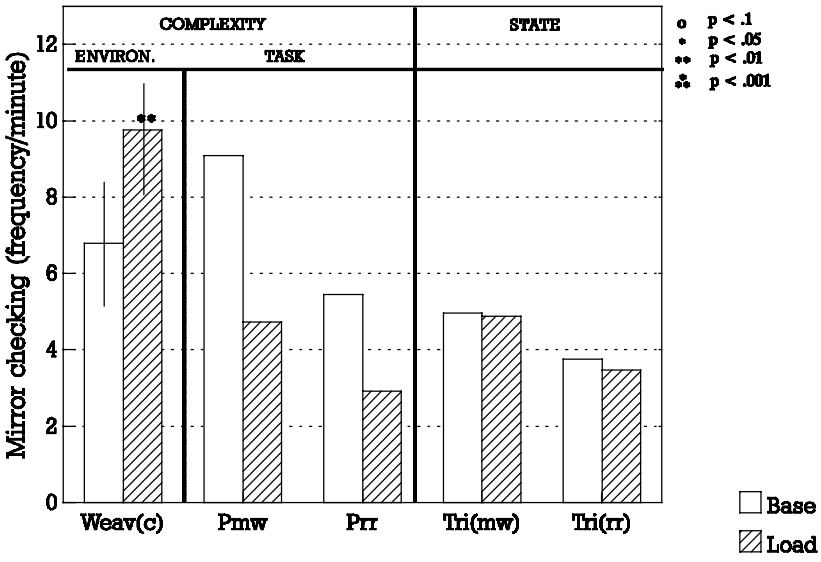


Figure 10. Frequency of interior and outside mirror checking in the car-phone and antihistamine study. In both studies motorway and ringroad sections were scored. If available the 95% confidence interval is indicated.

In the car-phone and antihistamine study, mirror checking *frequency* was scored from video, in both studies separately for the (quiet) motorway and (busy) ringroad. In the Weaving Section study the CEMRE-condition could be used to assess mirror-scanning frequency. As can be seen in figure 10, frequency of mirror-looking is reduced in the load condition of the car-phone study. The main effect of car-phone was not significant, but the interaction between road type and phone was. The larger effect of load on the motorway may be responsible for

this. No effect of load was found in the antihistamine study, only the effect of road type (again ringroad vs. motorway) was significant. In both studies mirror-looking frequency was lower on the more traffic-dense ringroad, where a car-following task had to be performed. In the Weaving Section study a significant *increase* in frequency of mirror checking was found in the load condition. This is particularly important because no difference in *duration*, i.e. proportion of the total time, of looking into the mirrors between the load and baseline conditions was found. Again the road environment may be responsible for the increase. The load section of the motorway was a combined entrance/exit with vehicles merging in and out of traffic, while the control section did not contain any entrances or exits. An increase in mirror-checking frequency and 'behind traffic monitoring' is important near entrances, even if no change-of-lane is planned, owing to the possible need of an evasive manoeuvre to the left-hand lane.

Rear mirror checking was also affected in the study reported by Van Winsum et al. (1989). In an unfamiliar environment, frequency of looking into the rear view mirror was reduced in the higher workload condition. Frequency of fixations seems most useful for workload assessment, though only if workload demand is not low. Fixation duration may be useful to assess certain aspects of task difficulty, in particular legibility, layout and amount of information (Fairclough et al., 1993).

Additional tasks

An actual additional task that had to be performed simultaneously to driving was the *PASAT*, the *Paced Serial Addition Task* (Gronwall & Sampson, 1974). The task itself is a demanding combination of a memory load and an addition test. This secondary task was used in the car-phone study, where the stimuli (digits) were presented over the phone. The task was used to create a fixed, heavy information-processing load on the subjects, more or less comparable to a difficult conversation. There was no control condition in which the task was performed without having to drive a car and/or use the car-phone. No significant differences in performance between the two road classes, motorway and ringroad, were found.

Earlier, at the end of the 60's, Brown et al. (1969) had studied the effects of telephoning on car driving performance by having subjects drive a car and perform gap-acceptance tests which were combined with a reasoning test. Subjects had to judge the correctness of sentences in relation to pairs of letters, e.g. "A follows B, -BA" (answer: True). Any impairment in driving performance could be attributed to divided attention; there was no need for the subjects to manually operate the car-phone. No effects of the additional task on primary-task vehicle-control measures were found, with the exception of an increase in time that was required to complete the circuit. Performance on the secondary task, however, was poorer in the

condition in which the task was combined with driving. Both reaction time and the proportion of errors increased. Gap-acceptance performance was also reduced by the additional task.

Verwey (1993b) carried out an experiment in which 48 subjects drove an instrumented vehicle over rural and inner-city roads while as secondary task they performed a visual detection task or an auditory addition task. While driving, subjects were guided by vocal messages issued by the experimenter. The experiment was a between-subject study with as factors: age (young vs. old), secondary task (auditory addition task vs. visual detection task), route familiarity (2 levels) and traffic density (2 levels). Subjects were instructed to give priority to the primary task of driving (Subsidiary Task Paradigm). Single-task performance of the secondary task while standing still was poorest for the elderly (79% opposed to 88% correct for the young). When driving, the older subjects' secondary-task performance (73% correct) was affected, while the younger subjects' performance did not decline (87% correct). Familiarity and traffic density had little effect on performance, while large differences on secondary-task performance were found between road situations. Between similar situations, i.e. between comparable road characteristics, no differences on secondary-task performance were found. In the study primary-task performance was only measured by assessment of speed control. Since different road segments had different speed limits, conclusions regarding primary-task performance are restricted. However, subjects unfamiliar with the road drove slightly slower and may therefore have reduced workload by adapting primary-task performance.

Brouwer et al. (1991) and Van Wolffelaar et al. (1990) have used an elegant 'driving-simulation' task. It was not the task environment that was elegant, but the way in which the level of primary-task performance was adapted to individual capability. By individually adapting the level of single task performance they succeeded in obtaining an equal task *difficulty* for all subjects. The primary task was a compensatory lane-tracking task. Added to this task was a visual analysis task. Van Wolffelaar (1990) added a third task to these, subjects had to respond to visual stimuli presented in the periphery. Although the simulator and the tasks that were used are more similar to laboratory tasks than to actual driving, the advantage of equal single-task difficulty for all is that divided attention problems can be studied taking into account differences in individual capability and allocation strategy. Results show that elderly are less successful in dividing attention in dual-task performance.

Properties of secondary-task measures

If it is assumed that performance of a secondary task uses up 'spare capacity', then secondary-task measures could be performance measures that are sensitive in the A region. However, most secondary tasks interfere (to a varying extent) with primary-task performance and

task instruction alone cannot determine which task receives priority. Embedded tasks are regarded as the best secondary tasks. Even though it still is not certain what priority the embedded task receives, at least primary-task intrusion is low. In car driving, measurement of car-following performance and mirror checking can supply embedded task measures. Delay in car-following was found to be a sensitive measure in the sedative antihistamine, alcohol and car-phone conditions. Sensitivity of this measure can accordingly be expected in the D/A1 and A3/B regions of performance. Frequency of mirror checking was found to be sensitive in the Weaving Section study, while the measure also differed between motorway and ringroad-driving. This measure is sensitive in the A3/B regions, while the frequency was not affected in the antihistamine study, and sensitivity in the A1/D regions requires further examination. Duration of glances in the mirror was not sensitive in the Weaving Section study, and no conclusions with respect to sensitivity of this measure in regions of performance can be drawn.

Both delay in car-following and mirror checking can reflect performance at the manoeuvre level. Diagnosticity of the latter measure to visual demand is moderate to high (Fairclough et al., 1993). Delay in following a lead vehicle was found to be sensitive in car-following conditions in all studies and seems to be a sensitive and reliable measure. Mirror checking frequency showed a similar sensitivity in the motorway and ringroad conditions of the antihistamine and car-phone studies, and reliability is accordingly rated high. Primary-task intrusion when using embedded secondary tasks is low. However, when studying car-following behaviour and more or less natural variations from a lead car have to be followed, task priorities may become somewhat blurred. Primary-task intrusion and operator acceptance when registering mirror-checking behaviour depends upon measurement technique. The CEMRE is an intrusive device while video registrations made by small cameras can remain completely unnoticed by the subjects. Implementation requirements in terms of instrumentation and time/equipment required for analysis are high for all measures. An overview of secondary-task performance measures' properties is presented in table 9. Mirror checking measures are based on a limited number of studies, and were measured with different techniques.

Apart from quantification of task performance in measures such as the SDLP or the frequency of mirror scanning, task performance could be rated by an observer. This method is sometimes used, but suffers from other methodological problems, such as training of the experimenter. If applied correctly, and if observers are well trained, results could add to the previously discussed primary and secondary task-performance measures. Critical incidents, law violations and lateral position errors are measures of driving performance and have been used as such in task-performance assessment (e.g., Pohlmann & Traenkle, 1994). In particular, complex behaviour, such as the

occurrence of critical incidents, or behaviour in a complex driving environment can be easier, or more accurately, detected and judged by an observer than captured in a single performance measure.

Table 9. Summary of properties of secondary-task workload measures.

	Measures		
property	Delay in car-following	Mirror Checking Duration	Frequency
sensitivity (Region)	D-A1,A3-B	?	A3-B
diagnosticity	low - moderate	moderate (?)	mod.-high
selectivity	moderate (?)	?	moderate (?)
reliability	high	?	high
primary-task intrusion	low-moderate	low	low
implementation requirements	high	high	high
operator acceptance	high	high	high

5.4 Physiological measures

Heart rate measures, ECG

Heart rate measures have been, and still are, very popular as in-vehicle registered physiological measures. The attractiveness of ECG is obvious, electrodes are easy to attach and distortion by physical movements is limited with car drivers, who simply have no other choice than to remain seated while driving.

Although heart period is measured and used as input for statistical analyses, the more popular 'average heart rate' during baseline and load condition is shown in figure 11. Note that the load condition is compared with a (similar) *baseline condition*, and not with the rest measurement. Compared to rest, driving (in both baseline and load condition) significantly elevates heart rate in all conditions. In the Noise Barrier experiment, average heart rate decreased in the load condition, while in the simulator (Tut) and antihistamine studies no effects of load compared with baseline measurements were found. An increase in heart rate (or a decrease in heart period or IBI) was found in both conditions of the Weaving Section study, and as a result of telephone use. On the adapted road leading through the woods (Wr), HR marginally significantly increased. Low amounts of alcohol increased HR on the motorway (Alc(mw)), an effect that is in accordance with findings of Mascord et al. (1995). The active drug Triprolidine did not affect heart rate frequency significantly, but average HR was prominently decreased as a result of time-on-task (the fatigue or 'vigilance' condition as it is indicated in appendix 5). These effects can even be better seen in figure 12, where the difference in

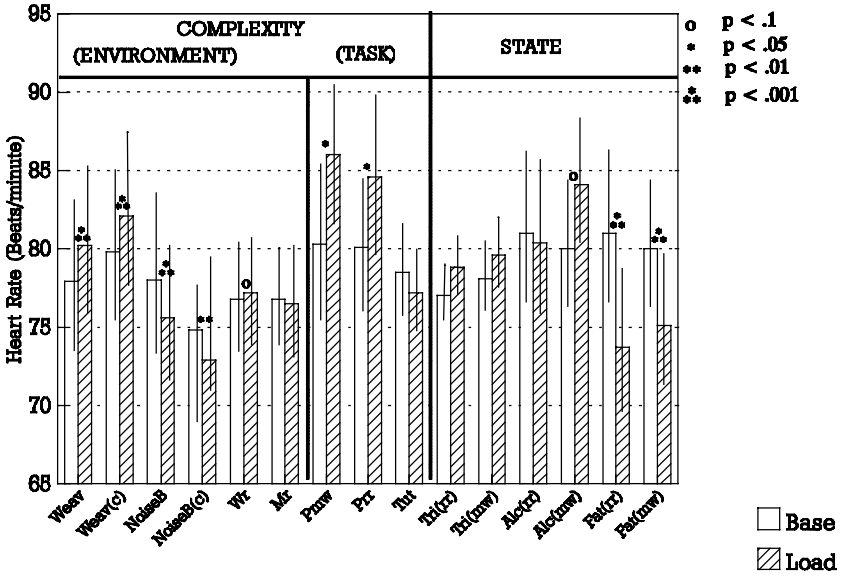


Figure 11. Average heart rate during baseline driving and during mental load. The 95% confidence interval is also indicated.

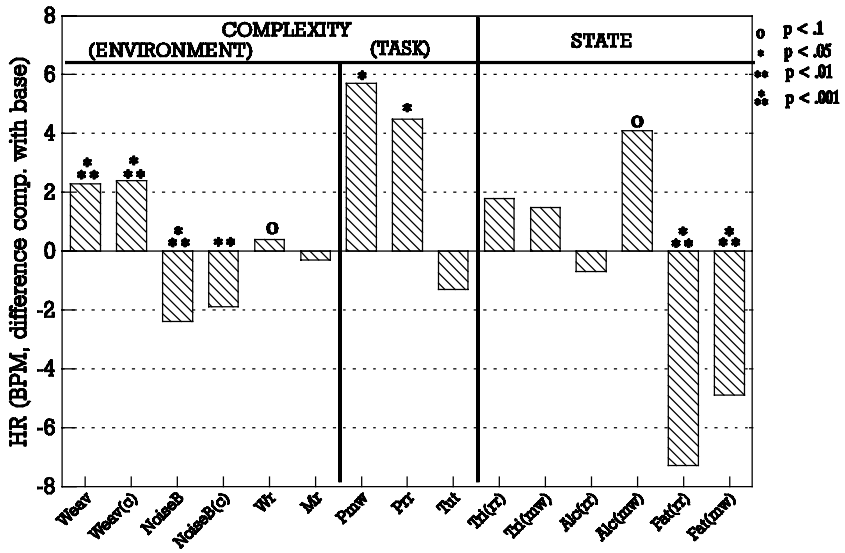


Figure 12. Difference in average heart rate during mental load compared with baseline driving.

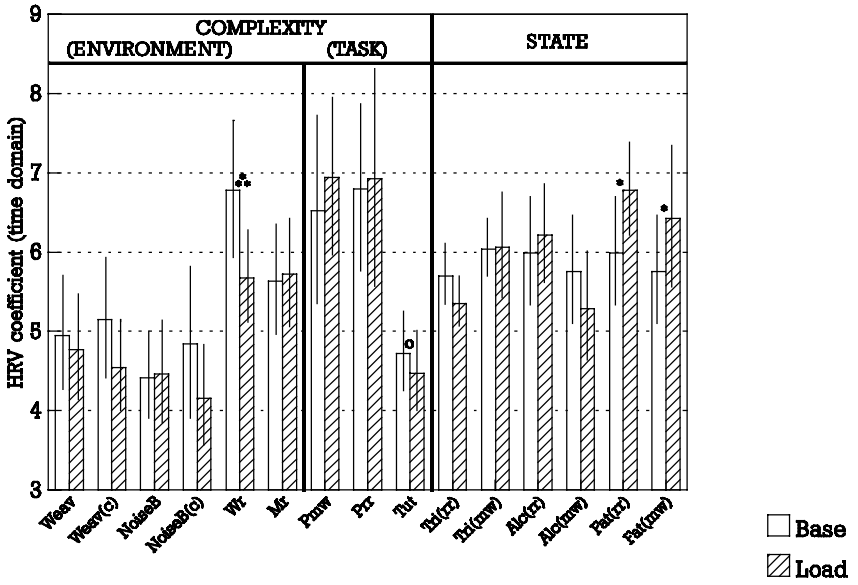


Figure 13. Standardized heart rate variability in the time domain during baseline driving and during mental load. The 95% confidence interval is also indicated.

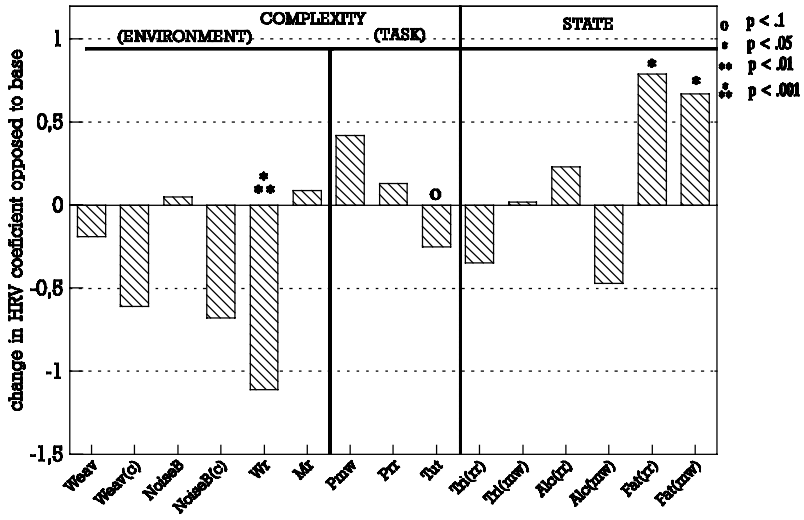


Figure 14. Difference in HRV during mental load compared with baseline driving.

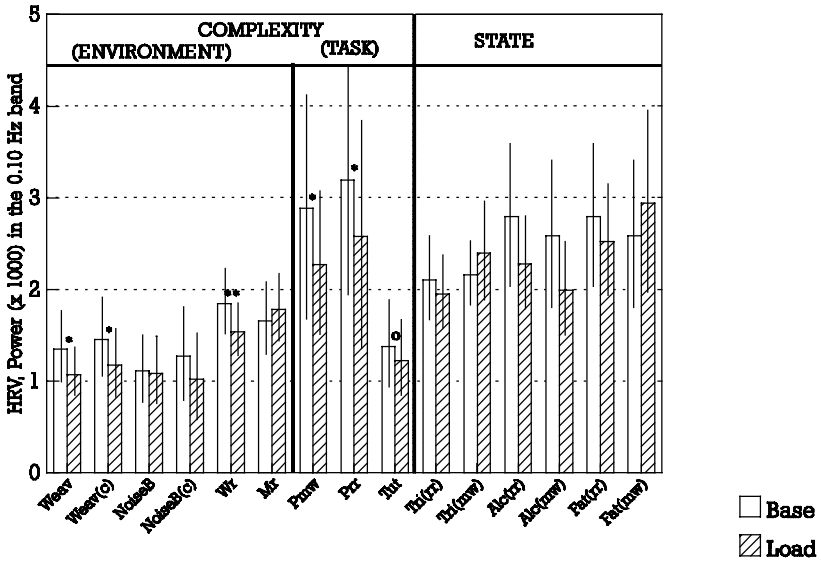


Figure 15. Energy in the 0.10 Hz frequency band of heart rate variability during baseline driving and during mental load. The 95% confidence interval is also indicated.

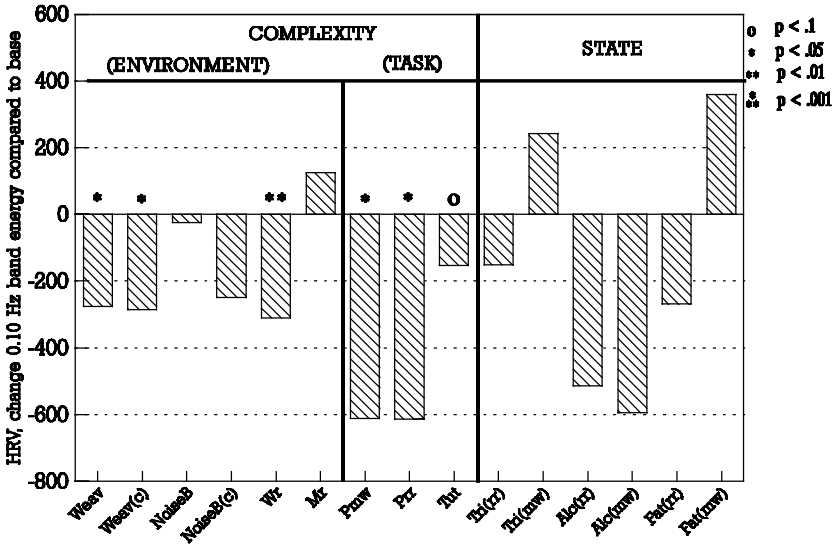


Figure 16. Difference in energy in the 0.10 Hz frequency band of heart rate variability during mental load compared with baseline driving

beats per minute of the load condition compared with the baseline condition are shown.

The Variation Coefficient (HRV), the standardized time-domain variability-measure of heart rate, is shown in figures 13 and 14. A significant decrease in variability was found in the DETER simulator study, and on the adapted Woodland road. A decrease was also found as a result of time-on-task; a finding in accord with Mascord & Heath (1992). The decrease in variability on the motorway as a result of an average Blood Alcohol Concentration of 0.5 ‰ was not statistically significant.

Compared to the time-domain variability measure, the frequency measure of 0.10 Hz variability is clearly more sensitive to the mental load manipulation (figures 15 and 16). Driving over the weaving section (Weav and Weav(c)), using the car-phone (Pmw and Prr), driving with feedback from the enforcement and tutoring system (Tut), as well as driving over the adapted road layout (Wr only) all reduced power in the 0.10 Hz variability band. The 0.10 Hz component-power is said to decrease as a result of relatively low levels of alcohol (see Gonzalez Gonzalez et al., 1992). The results regarding the Alcohol condition in the DREAM study (see figure 15) are in the expected direction, but not statistically significant.

Heart rate's idiosyncratic nature as well as high initial values can become very prominent in the spectral analysis and power computations that are required for determination of the 0.10 Hz HRV component. For this reason, energy in the 0.10 Hz frequency band is sometimes expressed as relative energy change compared with rest measurement (e.g., L.J.M.Mulder, 1988, Heino et al., 1996). For the studies in which a rest measurement was available, the additional change in 0.10 Hz HRV energy in the baseline and load conditions are displayed in table 10. In this table, the difference between baseline and load is also shown. Apart from 'size' differences, no large dissimilarities with figures 15 and 16 are apparent, with the exception of the Weaving Section and Noise Screen conditions in which the base-load difference is prominently reduced, or changes into a HRV increase in the non-CEMRE condition. The differences expressed as proportional change are, for reasons of lower inter-subject variability, probably more reliable than the absolute differences as shown in figure 16.

Heart rate profiles are a fairly recent development to monitor heart rate (variability) at a more continuous level. In the Weaving Section study (appendix 1), heart rate and 0.10 Hz-component heart rate variability were calculated and linked to specific road segments. Data chunks of 30 s were used as input and a resolution of 10 s was reached. With this technique, a more continuous index of the parameters can be obtained. In the Weaving Section study, changes in HR(V) during driving seem to reflect mental effort. Effects on parameters were tested

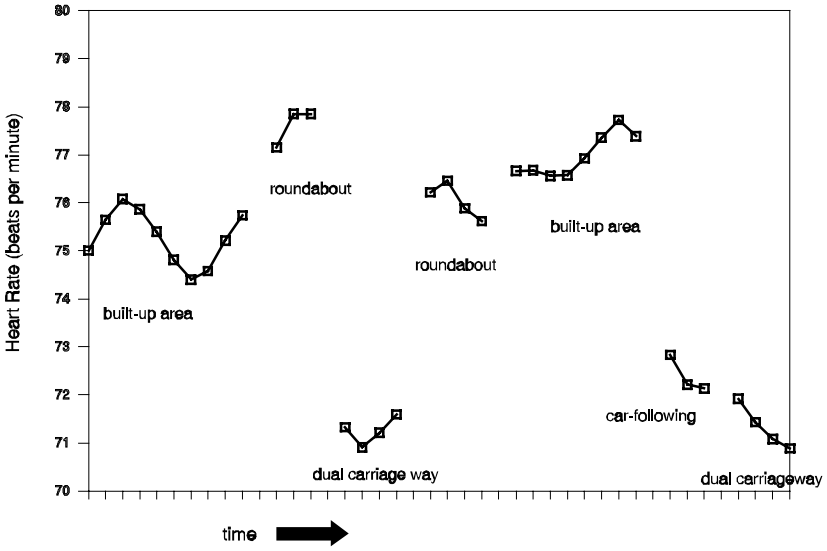


Figure 17. Average heart rate (N = 22) while driving in a simulator, during a trial in which subjects received feedback about detected law violations. Data were taken from the Tutoring (DETER) experiment (see appendix 4)

by comparing individual scores on an experimental section where load was suspected, with the scores on a section directly before this section (see appendix 1). The profile method was also applied in the simulator

Table 10. Change in energy in the 0.10 Hz frequency band of heart rate variability expressed as proportional change compared with rest measurements during baseline driving and during driving under mental load

study	Rest	Base	Load	Additional Change Load
complexity (environment)				
Weav	100%	-51%	-55%	- 4%
Weav(c)	100%	-30%	-53%	-23%
NoiseBarrier	100%	-18%	- 4%	+14%
NoiseBarrier(c)	100%	- 1%	-13%	-12%
Wr	100%	- 8%	-26%	-18%
Mr	100%	-19%	-11%	+ 8%
complexity (task)				
Pmw	100%	+13%	-14%	-27%
Prr	100%	+19%	-12%	-31%
Tut	100%	-13%	-23%	-10%

study and in figure 17 and 18 respectively, average heart rate and change in 0.10 Hz HRV energy are indicated; both averaged over 22 subjects. The third trial (see appendix 4) in which subjects received

feedback if violations were made, was selected for the figures. Thirty-second segments of data were used as input while the chosen step size again created a 10 second resolution. The different road environments are indicated in the figures. Clearly visible are the reductions in average heart rate frequency while driving over the dual carriageways and the increase in heart rate while driving around the roundabouts, and in the built-up areas. Figure 18 supports the idea that heart rate variability provides a reliable reflection of mental effort associated with different tasks. It can be seen that waiting for a red traffic light coincides with increases in variability, while driving on a roundabout corresponds to decreases in heart rate variability. The effects found in the simulator are very similar to effects found in an early on-the-road test of car driving, reported in Mulder (1980). Traffic density and traffic complexity were found to have a clear relation with reduced 0.10 Hz heart rate variability.

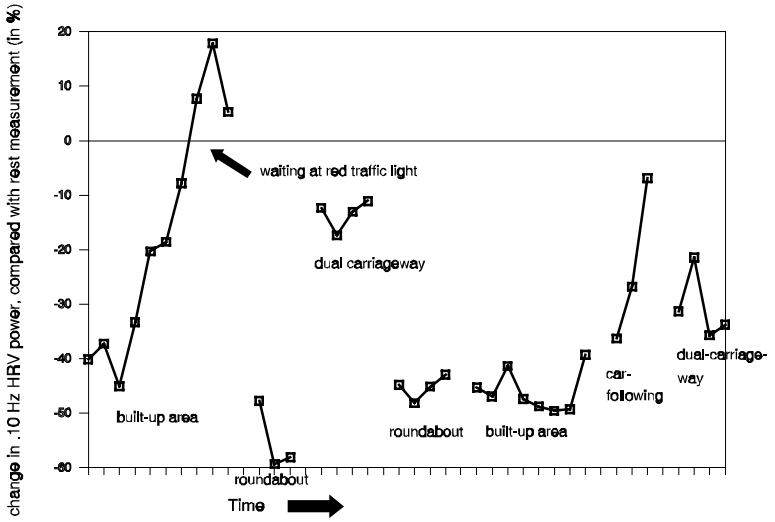


Figure 18. Change (in percentage) in 0.10 Hz HRV energy compared to the rest measurement. The same condition as in figure 17 was selected

a combined statistical test

Again the effects found in the different experiments were tested in combination. The overall effect of complexity on heart rate is a significant decrease in IBI (an increase in HR). The driver state test is largely dominated by the effect of fatigue in HR. The total effect of reduced driver state is a reduced heart rate. Due to the direct effect of alcohol on heart rate this result has to be regarded with caution. The test on heart rate variability in the time domain (the variation coefficient) shows that HRV is reduced under increased (environment) complexity, but not as a result of increased complexity due to additional tasks.

Inter-beat-intervals	
Complexity[b]:	z = -2.73, $p < 0.005$
Complexity (environment)[c]:	z = +0.20, NS
Complexity (task)[b]:	z = -3.34, $p < 0.0005$
State:	z = +1.73, $p < 0.05$
Heart rate variability (Time domain)	
Complexity[b]:	z = -1.95, $p < 0.05$
Complexity (environment)[c]:	z = -3.05, $p < 0.005$
Complexity (task)[b]:	z = +0.37, NS
State:	z = +0.65, NS
0.10 Hz component of heart rate variability	
Complexity[b]:	z = -3.23, $p < 0.0005$
Complexity (environment)[c]:	z = -2.53, $p < 0.01$
Complexity (task)[b]:	z = -2.29, $p < 0.025$
State:	z = -1.16, NS

[b] = Weighted, $\alpha_{Prr} = \alpha_{Pmw} = 1$, $\alpha_{Tut} = 2$.
[c] = All conditions equal weights

Only driver fatigue has a significant effect on HRV (increase), the total test of reduced driver state is not significant. Finally, spectral energy of heart rate variability in the 0.10 Hz frequency band is consistently and significantly reduced in the increased complexity conditions, but is not significantly affected in the test of the effect of a reduced driver state. This last aspect is very important and supports Mulder's idea (G. Mulder, 1995) that the 0.10 Hz component is sensitive to task-related effort and not to state-related effort.

Other physiological measures used

ElectroEncephalo Gram (EEG)

Ongoing EEG was more frequently used as indicator of driver state than as indicator of driver workload. The two are, however, not unrelated. As argued by some authors (Schneider et al., 1984, Kantowitz, 1992a), fatigue, e.g. as a result of the time spent performing a task, will be accompanied by a decreased arousal level and a reduced capacity, or a reduced willingness to spend resources (Meijman, 1991), and may therefore increase mental load. Ingestion of sedative drugs can be expected to result in the same effect. Brookhuis et al. (1985b, 1986) have found major increases in alpha and theta energy that were related to decreased driver activation caused by the use of antidepressant drugs. During prolonged train (Thorsvall & Åkerstedt, 1987) or truck driving (Kecklund & Åkerstedt, 1993), the driver's activation level as indicated by energy in the alpha and/or theta band was found to decrease rapidly. We (De Waard & Brookhuis, 1991a, appendix 5) have used the relative energy parameter [(alpha + theta) / beta] as indicator of driver state and

found a significant increase on the parameter with time-on-task. When, after two hours of non-stop driving, subjects returned to a busy ringroad and had to follow a lead car, activation level increased again. Clearly, the increased task demands on the ringroad, increased mental load. It seems that EEG frequency analysis are most useful as an indicator of tonic driver activation, and can be included in workload research for these purposes.

Electromyogram (EMG)

Facial EMG of the corrugator supercilii muscle was measured in the road-layout experiment (appendix 2). An effect of driving vs. rest, and of the two different road environments was found, while no effect of mental load as a result of the experimental road-layout was found. As the 0.10 Hz component of heart rate variability was sensitive to the (expected) difference in workload between the experimental and control road, and EMG activity of the corrugator was not, it is suggested that these measures may be tapping different dimensions of task load (see appendix 2). To my knowledge, no experimental field-studies that further examine the differential sensitivity to workload of these two measures have as yet been performed.

Eye movements

The number and duration of eye fixations on instruments or in the mirrors while driving (see for mirror scanning also the section on secondary-task performance) may well be indicative for driver strategy. Rockwell (1988) found more glances instead of longer glances at a radio that had to be adjusted while driving. The strategy for most of the complex tasks was to take a series of glances of 1.25 s until the task was completed. Only if information could not be extracted in a glance, e.g. due to legibility, drivers could be tempted to increase glance duration. A minority of glances of up to 3 s were found when adjusting the stereo. Rockwell (1988) argues that glances of this duration are a threat to traffic safety, in particular in car-following situations.

In the Noise Screen and Weaving Section studies, fixation time (as proportion of the total looking time) was determined for various categories. Parkes (1991) refers to this measure as ‘glance allocation’. In the studies initially eye movements were scored in various categories that were later combined into larger categories. Three categories were analyzed:

- traffic relevant fixations: looking straight forward, at other traffic, at the blind spot
- traffic irrelevant fixations: fixations on the other carriageway (which is irrelevant for motorway driving), the road environment, noise barriers, in the air
- mirrors & dashboard (‘other points of focus’)

The opportunity to look at, for driving, irrelevant stimuli will increase with decreases in workload (low time-pressure). This is partly

comparable to the path-neglect time in TLC (see under primary-task performance measures). A more demanding task *environment* requires an increase in the time spent looking at the road. In particular, the time spent looking at, for task performance relevant, objects, such as other traffic participants, road signs, road layout, etcetera, will increase. This includes looking in the mirrors. If it is not the road environment that requires additional attention but a device inside the car, it may have the opposite effect. Less time will be spent looking at relevant objects in the traffic environment.

In the Weaving Section study a reduction in time spent looking at the dashboard (speedometer) was found in the mental load condition (see table 11), while in the NoiseBarrier study, only data regarding the load condition were available. It is therefore difficult to draw conclusions on the basis of one study only. In addition to this, fixation time was scored in these analyses, and not fixation frequency, which is additionally required to asses driver strategy (Rockwell, 1988). In the Weaving Section study, fixation frequency on relevant objects increased in the load condition. While fixation time increases significantly with 6%, the number of fixations on traffic relevant objects is elevated with 13.2 fixations, an increase of 56%. Data from this study thus indicate larger sensitivity for fixation frequency compared with fixation duration expressed as proportion looking time. Scanning behaviour in which more glances instead of longer glances are taken (cf. Rockwell, 1988) could account for this difference in measure sensitivity.

Table 11: Proportion fixation time (%) and number of fixations per minute (fix/min) per category for the Weaving Section study (base and load) and the Noise Barrier study (load only). Eye movements were scored from video registrations made with the CEMRE equipment ('c' - condition only). Significant results have been printed in **bold**.

Study:	NoiseBarrier	Weaving Section		Weaving Section	
condition:	load	base	load	base	load
Fixation category	(%)	(%)	(%)	(fix/min)	(fix/min)
Relevant	76	72	**78**	23.4	**36.6**
Mirrors	3	10	11	6.8	**9.8**
Dashboard	7	6	**3**	4.8	4.7
Non-relevant	15	12	**8**	9.4	**6.9**

Results from other studies

ECG

In other studies found in the literature similar effects of mental load on ECG measures are reported as were found in the studies listed in table 3. Zeier (1979) measured heart rate in heavy city traffic while subjects drove a car with manual transmission, a car with automatic transmission or were just passengers. Both average heart rate and HRV (time domain) differed significantly between the manual-transmission condition and the other two conditions. Driving with automatic

transmission or riding as a passenger did not lead to a significant difference in heart rate measures.

Egelund (1982) concluded that the 0.10 Hz component was an indicator of driver fatigue. Although average HR decreased with time-on-task, Egelund found that HR was, just as time-domain-HRV, not sensitive to fatigue in this study. Janssen & Gaillard (1985) concluded that the 0.10 Hz component of heart rate variability was a more sensitive measure in mental load assessment than the P_{300} amplitude in ERPs in their on-the-road study.

Fairclough et al. (1991) found an effect on HR of car-phone use. Average heart rate while performing a secondary task presented through a hands-free phone was found to be higher compared with the same task presented by an experimenter that accompanied the driver in the passenger seat. The authors give two possible explanations for the effect, either additional effort is required in the phone condition due to lack of cues in conversation, or unfamiliarity with cellular mobile phones aroused the subjects (cf. the practice effects found by Brookhuis et al., 1991). Van Winsum et al. (1989) found an effect of mental load on average HR and on the 0.10 Hz component of HRV. They found navigation based on a map to be more effortful than navigation by vocal messages, as measured by a decrease in power in the 0.10 Hz component band of HRV.

Janssen et al. (1994) did not find significant effects on the 0.10 Hz component in an on-the-road study in which a control group and two groups that received driver support were compared. The trend in the displayed figure, however, indicated decreased variability with driver support, a situation that could be comparable to the DETER Tutoring study. The authors suggested that the measure's insensitivity could be due to sensitivity to 'an averaged workload level'. If so-called heart rate (variability) profiles had been determined, a more detailed picture might have emerged in that study.

EMG

One of the facial muscles that has been found to be sensitive to workload, is the frontalis (e.g., Van Boxtel & Jessurun, 1993). Zeier (1979) did not find an effect on EMG frontalis-activity of driving a car with automatic vs. manual gear transmission. However, he did find an effect of driving vs. being a passenger, the latter leading to lower muscle tension. The findings of Zeier (1979) support the idea that facial EMG activity taps a different dimension than (the 0.10 Hz component of) heart rate variability. Both in Zeier's study and the road layout study, EMG and HRV were differentially sensitive to workload. In addition, the two muscles that were measured, corrugator and frontalis, might also differ in selectivity. Jäncke (1994) found that the frontalis is not sensitive to emotional evaluation, while the corrugator is. A practical constraint of measurement of the corrugator in driving are the electrode positions that may interfere with the visual field.

ERPs

Measurement of Event Related Potentials (ERPs) has mainly been restricted to laboratory experiments. An exception to this are the studies reported by Janssen & Gaillard (1985). In two studies subjects had to drive an instrumented car through three road environments: through the city, over rural primary-roads and over motorways. During these rides they had to perform a secondary, auditory, Sternberg task. EEG was measured and P_{300} amplitude and its latency to task-relevant stimulus presentation (a secondary task) was determined. In the first experiment P_{300} amplitude was decreased and latency increased as a result of task load. City driving caused the largest increase in latency, surprisingly followed by motorway driving. In addition, motorway driving decreased P_{300} amplitude most, while amplitude was equally decreased during city and rural primary road driving, compared with rest measurements. In the second, similar study, city driving was left out. No effects on the P_{300} were measured in this experiment. The authors report large individual differences and significant variance in the ERP data. They relate the remarkable position of motorway driving compared with the other conditions to the self-pacedness of the driving task. Complexity of the selected motorway section may, however, have had an effect on task demands (e.g., driving of a clover-leaf was included).

EDA

In different studies Electrodermal Activity (EDA) has been related to the traffic environment (for an overview see Fairclough, 1993). Michaels (1962) reports an increase in EDR amplitude with an increase in traffic density, while Brown & Huffman (1972) report an increase in SCL if there is more traffic and there are more traffic lanes. Most in-vehicle studies have been performed in the sixties and focused on the effect of traffic environment on driver's EDA. In the seventies, Zeier (1979) measured EDA with electrodes positioned on the inner side of the left foot. He compared the effect of three conditions on psychophysiological measures, driving a car with manual transmission, with automatic transmission or being a passenger in a car. Effects on Skin Conductance level were not significant, but SCR (Skin Conductance Responses) were most numerous while driving the car with manual transmission. Least SCR were measured in the condition where subjects were passengers.

EDA is not only sensitive to all SNS activation, it might also be susceptible to physical movements. This last aspect is particularly relevant in car driving where EDA generally is measured on the palm of the hand, while both hands have to be used in steering. In mental workload research EDA might be useful to assess overall SNS activation level, but movements artifacts are a possible source of disturbance.

Hormones

There are not many mental load studies that include the evaluation of hormone levels. In general, the measurement of hormone

levels is restricted to situations in which the driver's occupation is very demanding. Examples of this type of stress research are the studies regarding city-bus drivers (Mulders et al., 1988) and coach drivers (Raggatt & Morrisay, submitted). One exception to the long-term impact studies was found, in a study reported by Zeier (1979) examining the effects of driving in heavy city traffic were examined. Adrenaline levels were found to be higher when driving a car as opposed to being a passenger. In addition, driving with manual transmission also led to higher adrenaline levels than driving with automatic transmission. No differences were found on noradrenaline levels.

properties of physiological measures

Background EEG is sensitive as an indicator of operator state, hence in region A to D. Average heart rate and heart rate variability in the time-domain are useful indicators of overall operator arousal level, i.e. in region D/B. The 0.10 Hz component however, is sensitive to task-related effort. It seems -as Mulder (1980) supposed- that the measure is sensitive to the Defense Response (Sokolov, 1963). The defense response is associated with a cardiovascular pattern of increased blood pressure, heart rate and stroke volume, decreased blood flow to renal, intestinal, and skin vascular beds, and increased skeletal muscle blood flow (Johnson & Anderson, 1990). The pattern is similar to responses evoked by stressful stimuli producing arousal in preparation for fighting. The defense response is coupled to increased sympathetic and reduced vagal activation, reflecting task-related effort and is accordingly connected to A3-region performance. Sensitivity of eye movements also seems to be highest in case of region A3 performance. Moreover, eye movements are related to visual demand, making it the highest diagnostic measure of table 12. Selectivity of EEG is low, operator state is reflected. The ECG measures differ in selectivity; HR and HRV are affected by many influences (respiration rate, physical effort) while this is less true for the 0.10 Hz component. Background EEG is a highly reliable, between-tests, measure for operator state, but individual differences (e.g., in the production of α-waves) weaken this qualification. The many tests in which ECG measures were found to be sensitive to workload result in a reliability that is rated high. Primary-task intrusion when taking EEG and ECG measures is low once the electrodes have been attached. Measurement of eye movements may interfere with primary-task performance if cornea reflection is registered with the aid of a CEMRE. Intrusion is low if the driver's face is registered on video or in case of registration of EOG. Implementation requirements are high for most physiological measures, as special equipment such as sensitive amplifiers are required. For spectral analysis, for example, precise, i.e. 1 ms resolution R-top detection is required (L.J.M.Mulder, 1992). Special software is also needed. Only when average heart rate and HRV are determined, are implementation

requirements less stringent. Finally, operator acceptance is inversely related to intrusiveness of measure registration. In table 12 the properties of different physiological measures are summarized.

Table 12. Summary of properties of physiological workload measures.

	Measures				
property	EEG back-ground	ECG HR	ECG HRV	ECG .10 Hz	Eye movements fixations/min
sensitivity (Region)	D-A2	D,B	D,B	A3	A3(?)
diagnosticity	low	low	low	low	high
selectivity	low	low	low-mod.	mod-high	?
reliability	mod-high	high	high	high	?
prim-task intrusion	low	low	low	low	[1]
implementation req.	high	moderate	moderate	high	high
operator acceptance	moderate	high	high	high	high-moderate[1]

[1] depends upon measurement technique

5.5 Discussion

Driving a vehicle is a task that demands continuous adaptation to a changing environment. A large part of the subtasks that have to be performed, such as lateral position control and speed maintenance, are tasks that are largely performed automatically at the control level, with hardly any driver effort. Representatives of performance measures at this level are the SDLP and steering wheel measures. At irregular intervals the control-level tasks are extended to include manoeuvre tasks, such as overtaking of other vehicles and following of leading cars. These tasks are not automated and require the driver's attention. Indicative measures of performance at this level are delay in car-following and the frequency of mirror checking.

A deteriorated driver state has been separated from increased task complexity as sources of increased workload. The effect of a deteriorated driver state and the increase in task complexity on primary-task performance might, however, appear to be the same. The primary-task parameter SDSTW changes in conditions of increased task complexity (e.g., Weaving Section study) and as a result of time-on-task. However, in combination with self-report ratings and physiology, a more differentiated picture emerges.

The pattern of measure sensitivity that emerges from the key studies (listed in table 3) is as follows: increased complexity, both in environment and in task, has an effect on the self-report scale RSME, and on the ECG. Task complexity vs. increases in environmental complexity seem to differentially affect the SDLP and SDSTW.

Additional tasks lead to a decrease in SDLP and SDSTW, while an increase in complexity of the environment increases both measures. An affected driver state resulting from the consumption of alcohol or sedative drugs does not affect heart rate variability as much as increases in complexity do. Time-on-task mainly affects the average heart rate level and the driver's EEG. Ratings on the self-report scale RSME and activation scale are more sensitive to changes in driver state. Secondary-task performance, in particular the embedded task of car-following, is sensitive to both sources of increased workload.

Region of performance remains a very important factor, as an increase in a primary-task parameter such as the SDLP can be the result of being overloaded as well as of driver deactivation. It seems that all deviations from optimal performance, both as a result of increased and decreased demand, can be traced by the combination of performance parameters and self-report and/or physiological indices. The moment task demands increase and the driver has to try harder, i.e. has to invest effort, heart rate variability in the 0.10 Hz band will decrease. The 0.10 Hz component is in particular sensitive to the defense response when task demands increase, and the driver exerts task-related effort. The changes on this parameter as a result of state-related effort and driver deactivation are less conclusive. Though the effects are large in terms of size, they fail to reach the 5% level of significance. Only Egelund (1982) reports significant changes on this parameter as a result of fatigue. The self-report scale RSME has more general sensitivity to driver effort, irrespective of whether it concerns state-related effort or task-related effort. It seems that these two measures, in combination with a primary-task performance measure, are the most useful to assess mental workload in the complete A region.

In most of the experiments listed in table 3, peak loads (Verwey & Veltman, 1995) play only a limited role. Workload during the car-phone conversation, while driving over the Weaving Section or over the adapted road layout; in all three conditions overall workload was increased. Only driving with the tutoring device could lead to peak loads at the moment messages are issued. However, on the basis of conversations with subjects after completion of the experiment it seems that the increase in mental workload in this experiment is more related to continuously intensified monitoring of the road environment and speedometer, than to information processing peaks at the moment of warnings. In sum, sensitivity of measures as reported above is sensitivity to overall workload, but no conclusions with respect to sensitivity to peak loads can be made on the basis of these experiments.

Task interpretation, goal setting

In the Car-phone study, Road layout and Tutoring experiments, an improvement on one of the primary-task measures, the SDLP, was found in the load condition. Since the effect of load in the three studies should be positioned in optimal performance section of the inverted-U, in the A3 region, no effect on primary-task parameters is expected. The

task environment may imply that higher performance is required and the improvement in performance may be the result of increased effort (as measured by a reduction in 0.10 Hz heart rate variability and an increased RSME score). In principle, the primary-task measure could therefore also be used for the assessment of workload in the A3 (and possibly also the A1) region. The best description of performance measures in these regions would then be 'no change or improvement in primary-task performance measures'. Finding an improvement in primary-task performance is paradoxical. Optimal performance is defined as the best performance, so no improvement is expected. In many laboratory tasks this is reasonable; in the field, however, conditions exist that allow for inaccuracies in primary-task performance during performance in the A-region. Unless subjects are given the strict instruction to drive in the centre of a lane and to try to steer as accurately as possible, improvement in primary-task performance can occur. A wide motorway lane, or the wide lanes used in the simulator experiment, do not necessitate accurate steering. *Goal setting or Task interpretation is an important factor and the need to perform at the highest level possible is in general absent in driving* and in field experiments. An improvement in performance was also found on the SDSTW-measure, in the load conditions of the Noise barrier and Tutoring studies. A similar explanation could be given for the improvement in lane-keeping performance, namely increased effort as indicated by physiological and self-report measures in both conditions results in increased primary-task performance.

Predicting the effects of tasks on driver mental workload is very difficult. Firstly, there are individual differences in goal setting and these differences vary from route choice to steering accuracy. Driving is to a large extent a self-paced task. If demands are too high, a slower driving speed can be chosen so as to be better able to deal with these demands. An elderly driver may prefer to make a detour so that he or she can drive over familiar roads thus facilitating the task environment. Once the task goals have been set, the task that has to be performed -the task demands- determine task complexity. How difficult a task is, however, depends upon capability (which may be lower for the elderly driver as just described), state and context. A novice driver will require more effort for vehicle control than an experienced driver. Driving performance itself can be related to externally set performance margins, critical levels, such as the margins proposed by Brookhuis (1995ab). Nevertheless only relative measures can give a further indication of mental workload. Strictly speaking, workload can only be determined per individual. It is always task X performed by individual Y (who is in a certain state) that leads to performance in Region Z. However, not all individuals are all that different and people often use similar strategies for performance of the same tasks. So, even though not all individuals set exactly the same goal, there are margins that are considered

acceptable. Heavy swerving and leaving the motorway lane is not considered acceptable by most drivers. Task demands can accordingly be defined in terms of maintaining the vehicle between the lines of the driving lane. For experienced young drivers it is not likely that there is much difference in (e.g. self-reported) effort required for the basic task of lateral and longitudinal vehicle control. This makes a link between a certain task and a region of task performance possible. In table 3 expectations about the region of performance for the different driving tasks have been specified.

Nevertheless, the most important factor in the measurement of workload is to assess changes in mental workload. Performance with the use of any device, in any environment or state under investigation, should be compared with baseline performance, driving without the use of the device, under 'normal' or standard conditions or while being sober. Changes in mental workload (measures) give a clear indication of what the effects of the changed demands are, incorporating at the same time changes in strategy or altered goals. This is, after all, the way people deal with changes in task demands in real life.

6
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

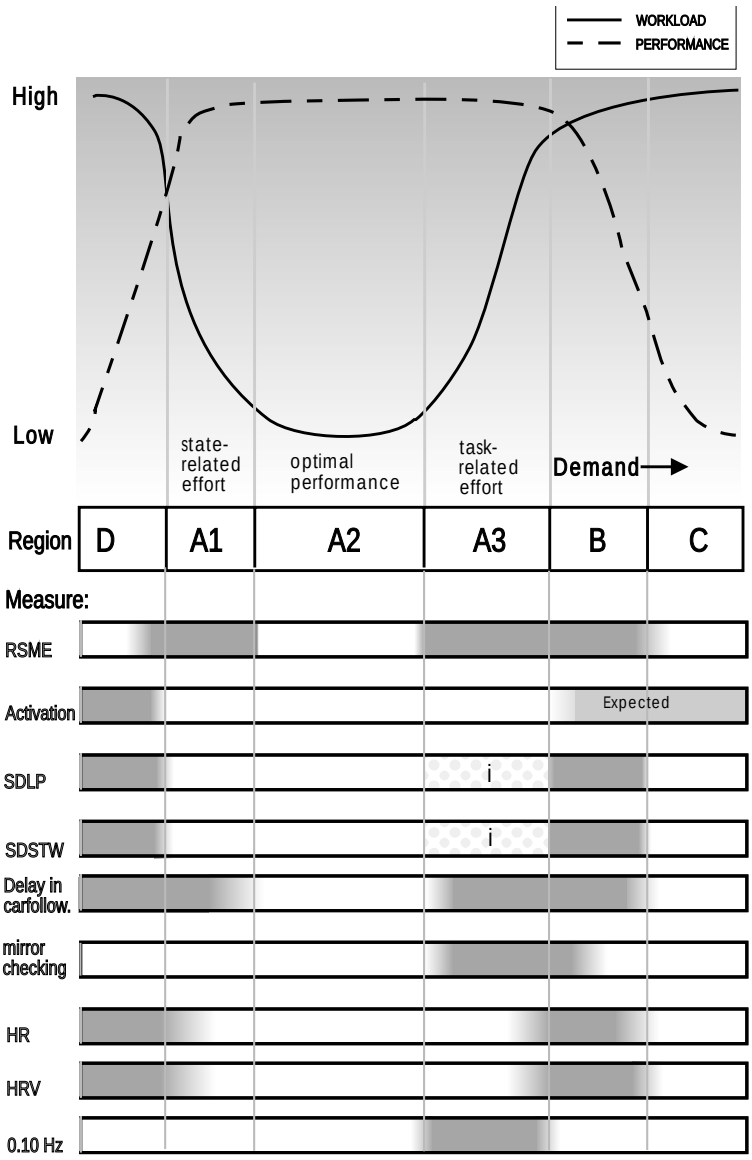


Figure 19. Workload in 6 regions and sensitivity of different measures to driver mental workload. RSME and Activation are self-report scales, SDLP and SDSTW are primary-task performance parameters. 'Delay carfoll.' is an embedded secondary task measure; the delay in following of speed changes of a lead 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.

7
References

Aasman, J., Mulder, G. & Mulder, L.J.M. (1987). Operator effort and the measurement of heart-rate variability. *Human Factors, 29*, 161-170.

Alm, H. & Nilsson, L. (1994). Changes in driver behaviour as a function of handsfree mobile phones-A simulator study. *Accident Analysis and Prevention, 26*, 441-451.

Backs, R.W. & Ryan, A.M. (1992). Multimodal measures of mental workload during dual-task performance: energetic demands of cognitive processes. In *Proceedings of the Human Factors Society 36th annual meeting* (pp. 1413-1417). Santa Monica, CA: Human Factors Society.

Backs, R.W. & Walrath, L.C. (1992). Eye movement and pupillary response indices of mental workload during visual search of symbolic displays. *Applied Ergonomics, 23*, 243-254.

Backs, R.W. & Seljos, K.A. (1994). Metabolic and cardiorespiratory measures of mental effort: the effects of level of difficulty in a working memory task. *International Journal of Psychophysiology, 16*, 57-68.

Bartenwerfer, H. (1969). Einige praktische Konsequenzen aus der Aktivierungstheorie. *Zeitschrift für experimentelle und angewandte Psychologie, 16*, 195-222.

Bauer, L.O., Goldstein, R. & Stern, J.A. (1987). Effects of information-processing demands on physiological response patterns. *Human Factors, 29*, 213-234.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin, 91*, 276-292.

Berntson, G.G., Cacioppo, J.T., Quigley, K.S. & Fabro, V.T. (1994). Autonomic space and psychophysiological response. *Psycho-physiology, 31*, 44-61.

Byers, J.C., Bittner, A.C. & Hill, S.G. (1989). Traditional and raw task load index (TLX) correlations: are paired comparisons necessary? In A. Mital (Ed.), *Advances in industrial ergonomics and safety, I* (pp 481-485). London: Taylor & Francis.

Blaauw, G.J. (1984). *Car driving as a supervisory control task*. PhD Thesis. TNO-Institute for Perception, Soesterberg, The Netherlands.

Brenner, M, Doherthy, E.T. & Shipp, T. (1994). Speech measures indicating workload demand. *Aviation, space, and environmental medicine, 65*, 21-26.

Broadbent, D.E. (1958). *Perception and communication*. London: Pergamon.

Brookhuis, K.A. (1989). *Event related potentials and information processing*. PhD Thesis. Groningen: University of Groningen.

Brookhuis, K.A. (1995a). *DETER, Detection, Enforcement and Tutoring for Error Reduction. Final Report*. (Report 2009/DETER/ Deliverable 20). Haren, The Netherlands: Traffic Research Centre, University of Groningen.

Brookhuis, K.A. (1995b). Driver Impairment Monitoring System. In M. Vallet & S. Khardi, *Vigilance et Transports. Aspects fondamentaux, dégradation et préventation*. Lyon, France: Presses Universitaires de Lyon.

Brookhuis, K.A., De Vries, G., Prins van Wijngaarden, P., Veenstra, G., Hommes, M., Louwerens, J.W. & O'Hanlon, J.F. (1985a). *The effects of increasing doses of Meptazinol (100, 200, 400 mg) and Glafenine (200 mg) on driving performance* (Report VK 85-16). Haren, The Netherlands: Traffic Research Centre, University of Groningen.

Brookhuis, K.A., Louwerens, J.W. & O'Hanlon, J.F. (1985b). The effect of several antidepressants on EEG and performance in a prolonged car driving task. In W.P. Koella, E. Rüther & H. Schulz (Eds.) *Sleep '84*, (pp. 129-131). Stuttgart: Gustav Fischer Verlag.

Brookhuis, K.A., Louwerens, J.W. & O'Hanlon, J.F. (1986). EEG energy-density spectra and driving performance under the influence of some anti-depressant drugs. In J.F. O'Hanlon & J.J. de Gier (Eds.), *Drugs and Driving* (pp. 213-221). London: Taylor & Francis.

Brookhuis, K.A., De Vries, G. & De Waard, D. (1989). *De effecten van het gebruik van de autotelefoon op het rijgedrag (The effects of mobile telephoning on driving performance*). (Report VK 89-02). Haren, The Netherlands: Traffic Research Centre, University of Groningen.

Brookhuis, K.A., De Vries, G. & De Waard, D. (1991). The effects of mobile telephoning on driving performance. *Accident Analysis and Prevention, 23*, 309-316.

Brookhuis, K.A., De Vries, G. & De Waard, D. (1993). Acute and subchronic effects of the H_1-histamine receptor antagonist ebastine in 10, 20 and 30 mg dose, and triprolidine 10 mg on car driving performance. *British Journal of Clinical Pharmacology, 36*, 67-70.

Brookhuis, K.A. & De Waard, D. (1993). The use of psychophysiology to assess driver status. *Ergonomics, 36*, 1099-1110.

Brookhuis, K.A., De Waard, D. & Mulder, L.J.M. (1994). Measuring driving performance by car-following in traffic. *Ergonomics, 37*, 427-434.

Brouwer, W.H., Waterink, W., Van Wolffelaar, P.C. & Rothengatter, J.A. (1990). Divided attention in experienced young and older drivers: lane tracking and visual analysis in a dynamic driving simulator. *Human Factors, 33*, 573-582.

Brouwer, W.H. & Ponds, R.W.H.M. (1994). Driving competence in older persons. *Disability and Rehabilitation, 16*, 149-161.

Brown, I.D. & Poulton, E.C. (1961). Measuring the spare 'mental' capacity of cardrivers by a subsidiary task. *Ergonomics, 4*, 35-40.

Brown, I.D., Thickner, A.H. & Simmonds, D.C.V. (1969). Interference between concurrent tasks of driving and telephoning. *Journal of Applied Psychology, 53*, 419-424.

Brown, J.D. & Huffman, W.J. (1972). Psychophysiological measures of drivers under actual driving conditions. *Journal of Safety Research, 4*, 172-178.

Cnossen, F. (1994). *Mental effort from energetical and computational perspectives*. Unpublished paper for the BCN course Cognitive Neuroscience. Haren, The Netherlands: Traffic Research Centre, University of Groningen.

Cooper, R., Osselton, J.W. & Shaw, J.C. (1980). *EEG Technology*. London: Butterworths

Dawson, M.E., Schell, A.M. & Filion, D.L. (1990). The electrodermal system. In J.T. Cacioppo & L.G. Tassinary (Eds.), *Principles of psychophysiology* (pp. 295-324). Cambridge: Cambridge University Press.

De Vries, G., De Waard, D. & Brookhuis, K.A. (1989). *A double blind study to compare the acute and subchronic effects of Ebastine 10, 20 and 30 mg o.d., triprolidine 10 mg o.d. and placebo on car driving performance* (Report VK 89-22). Haren, The Netherlands: Traffic Research Centre, University of Groningen.

De Waard, D. (1991). Driving behaviour on a high-accident-rate motorway in the Netherlands. In C. Weikert, K.A. Brookhuis and S. Ovinius (Eds.), *Man in complex systems*, Proceedings of the Europe Chapter of the Human Factors Society Annual Meeting. Work Science Bulletin 7 (pp 113-123). Lund, Sweden: Work Science Division, Department of Psychology, Lund University.

De Waard, D., Van der Linden, L.H.K. & Westra, E.J. (1990). *Observaties van rijgedrag in een geïnstrumenteerde auto op de A-28 tussen Assen en Groningen (Observations of driving behaviour in an instrumented vehicle on the A28 motorway between Assen and Groningen).* (Report VK 90-16). Haren, The Netherlands: Traffic Research Centre, University of Groningen.

De Waard, D. & Brookhuis, K.A. (1991a). Assessing driver status: a demonstration experiment on the road. *Accident Analysis and Prevention, 23*, 297-307.

De Waard, D. & Brookhuis, K.A. (1991b). The feasibility of a device that monitors driver's status and abilities. In *Proceedings of the 24th ISATA International symposium on automotive technology and automation* (pp. 725-731). Croydon, England: Automotive Automation Limited.

De Waard, D., Brookhuis, K.A., Van der Hulst, M. & Van der Laan, J.D. (1994). *Behaviour comparator prototype test in a driving*

simulator (Report 2009/DETER/Deliverable 10 (321B)). Haren, The Netherlands: Traffic Research Centre, University of Groningen.

De Waard, D., Jessurun, M., Steyvers, F.J.J.M., Raggatt, P.T.F., Brookhuis, K.A. (1995). Effect of road layout and road environment on driving performance, drivers' physiology and road appreciation. *Ergonomics, 38*, 1395-1407.

De Waard, D. & Steyvers, F.J.J.M. (1995). *Wegbelijning in rurale verblijfgebieden: een experiment met kantbelijning (Road delineation in rural residential areas: an experiment with edgelines).* (Report VK 95-07). Haren, The Netherlands: Traffic Research Centre, University of Groningen.

De Waard, D. & Brookhuis, K.A. (in press). Behavioural adaptation of drivers to warning and tutoring messages. Results from an on-the-road and simulator test. Special Issue 'Driver Vehicle Interaction', *International Journal of Vehicle Design*.

De Waard, D., Van der Hulst, M. & Brookhuis, K.A. (submitted). Elderly and young driver's reaction to an in-car enforcement and tutoring system.

Dimberg, U. (1988). Facial electromyography and the experience of emotion. *Journal of Psychophysiology, 2*, 277-282.

Dimberg, U. & Thell, S. (1988). Facial electromyography, fear relevance and the experience of stimuli. *Journal of Psychophysiology, 2*, 213-219.

Egelund, N (1982). Spectral analysis of heart rate variability as an indicator of driver fatigue. *Ergonomics, 25*, 663-672.

Eggemeier, F.T. & Wilson, G.F. (1991). Performance-based and subjective assessment of workload in multi-task environments. In D.L. Damos (Ed.), *Multiple-task performance*. (pp. 217-278). London: Taylor & Francis.

Eggemeier, F.T., Wilson, G.F., Kramer, A.F. & Damos, D.L. (1991). Workload assessment in multi-task environments. In D.L. Damos (Ed.)., *Multiple-task performance*. (pp. 207-216). London: Taylor & Francis.

Fairclough, S.H. (1991). *Adapting the TLX to measure driver mental workload* (Report V1017/BERTIE/No. 71). Loughborough, Leics, UK: HUSAT Research Institute.

Fairclough, S. H. (1993). Psychophysiological measures of workload and stress. In A.M. Parkes & S. Franzèn (Eds.), *Driving Future Vehicles*. (pp. 377-390). London: Taylor & Francis.

Fairclough, S.H. (Ed.) (1994). *Driver State Monitor* (Report V2009/DETER/Deliverable 5 (330A)). Haren, The Netherlands: Traffic Research Centre, University of Groningen.

Fairclough, S.H., Ashby, M.C., Ross, T. & Parkes, A. (1991). Effects of handsfree telephone use on driving behaviour. In *Proceedings of the 24th ISATA International symposium on automotive technology and automation* (pp. 403-409). Croydon, England: Automotive Automation Limited.

Fairclough, S.H., Ashby, M.C. & Parkes, A.M. (1993). In-vehicle displays, visual workload and usability evaluation. In A.G. Gale, I.D. Brown, C.M. Haslegrave, H.W. Kruysse & S.P. Taylor (Eds.), *Vision in vehicles -IV* (pp. 245-254). Amsterdam: North-Holland.

Fibiger, W., Christensen, F., Singer, G. & Kaufmann, H. (1986). Mental and physical components of sawmill operatives' workload. *Ergonomics, 29*, 363-375.

Frankenhaeuser, M. (1989). A biopsychosocial approach to work life issues. *International Journal of Health Services, 19*, 747-758.

Fridlund, A.J. & Cacioppo, J.T. (1986). Guidelines for human electromyographic research. *Psychophysiology, 23*, 567-589.

Godthelp, J. (1984), *Studies on human vehicle control*. PhD Thesis, Soesterberg, The Netherlands: Institute for Perception, TNO.

Godthelp, J. (1988). The limits of path error-neglecting in straight lane driving. *Ergonomics, 31*, 609-619.

Godthelp, J., Milgram, P. & Blaauw, G.J. (1984). The development of a time-related measure to describe driving strategy. *Human Factors, 26*, 257-268.

Gonzalez Gonzalez, J., Mendez Llorens, A., Mendez Novoa, A. & Cordero Valeriano, J.J. (1992). Effect of acute alcohol ingestion on short-term heart rate fluctuations. *Journal of Studies on Alcohol, 53*, 86-90.

Gopher, D. & Braune, R. (1984). On the psychophysics of workload: Why bother with subjective measures? *Human Factors, 26*, 519-532.

Gopher, D. & Sanders, A.F. (1984). S-Oh-R: Oh stages! Oh resources!. In W. Prinz & A.F. Sanders, *cognition and motor processes*. (pp 231-253). Berlin: Springer Verlag.

Gopher, D. & Donchin, E. (1986). Workload -an examination of the concept. In K.R. Boff, L. Kaufman & J.P. Thomas (Eds.), *Handbook of perception and human performance. Volume II, cognitive processes and performance*. (pp 41/1-41/49). New York: Wiley.

Green, P., Williams, M., Hoekstra, E, George, K. & Wen, C. (1993a). *Initial on-the-road tests of driver information system interfaces: route guidance, traffic information, IVSAWS, and vehicle monitoring* (Report UMTRI-93-32). Ann Arbor, MI, U.S.A.: The University of Michigan Transportation Research Institute.

Green, P., Lin, B. & Bagian, T. (1993b). *Driver Workload as a function of road geometry: a pilot experiment* (Report UMTRI-93-39). Ann Arbor, MI, U.S.A.: The University of Michigan Transportation Research Institute.

Gronwall, D.M.A. & Sampson, H. (1974). *The psychological effects of concussion*. Auckland, New Zealand: Auckland University Press.

Grossman, P. (1992). Respiratory and cardiac rhythms as windows to central and autonomic biobehavioral regulation: selection of

window frames, keeping the panes clean and viewing the neural topography. *Biological Psychology, 34*, 131-161.

Hancock, P.A. & Parasuraman, R. (1992). Human Factors and safety in the design of Intelligent Vehicle-Highway Systems (IVHS). *Journal of Safety Research, 23*, 181-198.

Hart, S.G. & Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In P.A. Hancock & N. Meshkati (Eds.), *Human Mental Workload* (pp 139-183). Amsterdam: North-Holland.

Hebb, D.O. (1955). Drives and the C.N.S. (Conceptual Nervous System). *Psychological Review, 62*, 243-254.

Hedman, L.R. & Sirevaag, E.J. (1991). In C. Weikert, K.A. Brookhuis and S. Ovinius (Eds.), *Man in complex systems*, Proceedings of the Europe Chapter of the Human Factors Society Annual Meeting. Work Science Bulletin 7 (pp 12-18). Lund, Sweden: Work Science Division, Department of Psychology, Lund University.

Heino, A., Van der Molen, H.H. & Wilde, G.J.S. (1990). *Risk-homeostatic processes in car following behaviour: Electrodermal responses and verbal risk estimates as indicators of the perceived level of risk during a car-driving task* (Report VK 90-22). Haren, The Netherlands: Traffic Research Centre, University of Groningen.

Heino, A., Van der Molen, H.H. & Wilde, G.J.S. (1996). Differences in risk experience between sensation avoiders and sensation seekers. *Personality and Individual Differences, 20*, 71-79.

Hendy, K.C., Hamilton, K.M. & Landry, L.N. (1993). Measuring subjective workload: when is one scale better than many? *Human Factors, 35*, 579-601.

Heslegrave, R.J., Ogilvie, J.C. & Furedy, J.J. (1979). Measuring baseline-treatment differences in heart rate variability: variance versus successive difference mean square and beats per minute versus interbeat intervals. *Psychophysiology, 16*, 151 -157.

Hicks, T.G. & Wierwille, W.W. (1979). Comparison of five mental workload assessment procedures in a moving-base driving simulator. *Human Factors, 21*, 129-143.

Hill, S.G., Iavecchia, H.P., Byers, J.C., Bittner, A.C., Zaklad, A.L. & Christ, R.E. (1992). Comparison of four subjective workload rating scales. *Human Factors, 34*, 429-439.

Hockey, G.R.J. (1986). A state control theory of adaptation and individual differences in stress management. In G.R.J. Hockey, A.W.K. Gaillard and M.G.H. Coles (Eds.), *Energetics and human information processing* (pp. 285-298). Dordrecht, The Nether-lands: Martinus Nijhoff Publishers.

Hoeks, L.T.M. (1995). *The pupillary response as a measure of mental processing load: with application to picture naming*. PhD Thesis, Nijmegen, The Netherlands: University of Nijmegen.

Hughes, P.K. & Cole, B.L. (1988). The effect of attentional demand on eye movement behaviour when driving. In A.G. Gale, M.H.

Freeman, C.M. Haslegrave, P. Smith & S.P. Taylor (Eds.), *Vision in vehicles-II* (pp. 221-230). Amsterdam: North-Holland.

Humphrey, D.G. & Kramer, A.F. (1994). Toward a psychophysiological assessment of dynamic changes in mental workload. *Human Factors, 36,* 3-26.

Hyndman, B.W. & Gregory, J.R. (1975). Spectral analysis of sinus arrhythmia during mental loading. *Ergonomics, 18,* 255-270.

Hyönä, J., Tommola, J. & Alaja, A.-M. (1995). Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. *The Quarterly Journal of Experimental Psychology, 48A,* 598-612.

Itoh, Y, Hayashi, Y., Tsukui, I. & Saito, S. (1990). The ergonomic evaluation of eye movement and mental workload in aircraft pilots. *Ergonomics, 33,* 719-733.

Jäncke, L. (1994). An EMG investigation of the coactivation of facial muscles during the presentation of affect-laden stimuli. *Journal of Psychophysiology, 8,* 1-10.

Janssen, W.H. (1979). *Routeplanning en -geleiding: een literatuurstudie (Route planning and route guidance: a review of the literature).* (Report IZF 1979-C13). Soesterberg, The Netherlands: Instituut voor Zintuigfysiologie.

Janssen, W.H. & Gaillard, A.W.K. (1985). EEG and heart rate correlates of task load in car driving. In A. Gundel (Ed.), *Proceedings of the workshop 'Electroencephalography in transport operations.* Köln, Germany: DFVLR-Institut für Flugmedizin.

Janssen, W.H., Kuiken, M.J. & Verwey, W.B. (1994). Evaluation studies of a prototype intelligent vehicle. In ERTICO (Ed.) *Towards an intelligent transport system. Proceedings of the first world congress on applications of transport telematics and intelligent vehicle-highway systems* (pp. 2063-2070). Boston: Artech House.

Jennings, J.R., Stringfellow, J.C. & Graham, M. (1974). A comparison of the statistical distributions of beat-by-beat heart rate and heart period. *Psychophysiology, 11,* 207-210.

Jessurun, M., Steyvers, F.J.J.M., De Waard, D., Dekker, K., Brookhuis, K.A. (1990). *Beleving, waarneming en activatie tijdens het rijden over een deel van de A2 (Appreciation, perception and activation while driving over a section of the A2 motorway).* (Report VK 90-18). Haren, The Netherlands: Traffic Research Centre, University of Groningen.

Jessurun, M., De Waard, D., Raggatt, P.T.F., Steyvers, F.J.J.M. & Brookhuis, K.A. (1993). *Implementatie van snelheidsbeperkende maatregelen op 80 km/uur wegen: effecten op rijgedrag, activatie en beleving (Implementation of speed-reducing measures on A-class roads: effects on driving performance, activation and appreciation).* (Report VK 93-01). Haren, The Netherlands: Traffic Research Centre, University of Groningen.

Johnson, A.K. & Anderson, E.A. (1990). Stress and arousal. In J.T. Cacioppo & L.G. Tassinary. *Principles of psychophysiology* (pp. 216-252). Cambridge: Cambridge University Press.

Jordan, P.W. & Johnson, G.I. (1993). Exploring mental workload via TLX: the case of operating a car stereo whilst driving. In A.G. Gale, I.D. Brown, C.M. Haslegrave, H.W. Kruysse & S.P. Taylor (Eds.), *Vision in Vehicles-IV*. (pp. 255-262). Amsterdam: North-Holland.

Jorna, P.G.A.M. (1992). Spectral analysis of heart rate and psychological state: a review of its validity as a workload index. *Biological Psychology, 34*, 237-257.

Jorna, P.G.A.M. (1993). Heart rate and workload variations in actual and simulated flight. *Ergonomics, 36*, 1043-1054.

Kahneman, D. (1973). *Attention and effort*. New Jersey, U.S.A.: Prentice Hall.

Kalsbeek, J.W.H. & Ettema, J.H. (1963). Scored regularity of the heart rate and the measurement of perceptual load. *Ergonomics, 6*, 306.

Kantowitz, B.H. (1987). Mental workload. In P.A. Hancock (Ed.), *Human Factors Psychology*. (pp 81-121). Amsterdam: North-Holland.

Kantowitz, B.H. (1992a). Heavy vehicle driver workload assessment: lessons from aviation. In *Proceedings of the Human Factors Society 36th annual meeting* (pp. 1113-1117). Santa Monica, CA: Human Factors Society.

Kantowitz, B.H. (1992b). Selecting measures for human factors research. *Human Factors, 34*, 387-398.

Kantowitz, B.H. & Knight, J.L. (1976). Testing tapping timesharing, II: auditory secondary task. *Acta Psychologica, 40*, 340-362.

Kecklund, G. & Åkerstedt, T. (1993). Sleepiness in long distance truck driving: an ambulatory EEG study of night driving. *Ergonomics, 36*, 1007-1017.

Korteling, J.E. (1994a). *Multiple-task performance and aging*. PhD Thesis. Groningen: University of Groningen.

Korteling, J.E. (1994b). Effects of aging, skill modification, and demand alternation on multiple-task performance. *Human Factors, 36*, 27-43.

Kramer, A.F. (1991). Physiological metrics of mental workload: a review of recent progress. In D.L. Damos (Ed.), *Multiple-task performance*. (pp. 279-328). London: Taylor & Francis.

Lee, D.H. & Park, K.S. (1990). Multivariate analysis of mental and physical load components in sinus arrhythmia scores. *Ergonomics, 33*, 35-47.

Louwerens, J.W., Brookhuis, K.A. & O'Hanlon, J.F. (1983). *The effects of the antidepressants Oxaprotiline, Mianserin, Amitryptiline and Doxepin upon actual driving performance* (Report VK 83-05). Haren, The Netherlands: Traffic Research Centre, University of Groningen.

Louwerens, J.W., Gloerich, A.B.M., De Vries, G., Brookhuis, K.A. & O'Hanlon, J.F. (1987). The relationship between drivers' blood alcohol concentration (BAC) and actual driving performance during high speed travel. In P.C. Noordzij & R. Roszbach (Eds.), *Alcohol, Drugs and Traffic Safety-T86* (pp. 183-186). Amsterdam: Excerpta Medica.

Macdonald, W.A. & Hoffmann, E.R. (1980). Review of relationship between steering wheel reversal rate and driving task demand. *Human Factors, 22*, 733-739.

Mascord, D.J. & Heath, R.A. (1992). Behavioral and physiological indices of fatigue in a visual tracking task. *Journal of Safety Research, 23*, 19-25.

Mascord, D.J., Walls, J. & Starmer, G.A. (1995). Fatigue and alcohol: interactive effects on human performance in driving-related tasks. In L. Hartley (Ed.), *Fatigue and Driving. Driver Impairment, Driver Fatigue and Driving Simulation* (pp.189-205). London: Taylor & Francis.

Matsumoto, R., Walker, B.B., Walker, J.M. & Hughes, H.C. (1990). Fundamentals of Neuroscience. In J.T. Cacioppo & L.G. Tassinary (Eds.), *Principles of psychophysiology* (pp. 58-112). Cambridge: Cambridge University Press.

May, J.G., Kennedy, R.S., Williams, M.C., Dunlap, W.P. & Brannan, J.R. (1990). Eye movement indices of mental workload. *Acta Psychologica, 75*, 75-89.

McLean, J.R. & Hoffmann, E.R. (1971). Analysis of drivers' control movements. *Human Factors, 13*, 407-418.

McLean, J.R. & Hoffmann, E.R. (1975). Steering reversals as a measure of driver performance and steering task difficulty. *Human Factors, 17*, 248-256.

Meijman, T.F. (1989). *Mentale belasting en werkstress. Een arbeidspsychologische benadering (Mental workload and workstress. A workpsychological approach)*. Assen, The Netherlands: van Gorcum.

Meijman, T.F. (1991) *Over vermoeidheid, arbeidspsychologische studies naar de beleving van belastingseffecten (About being tired, workpsychological studies into the perception of the effects of demand)*. PhD thesis, Rijksuniversiteit Groningen. Amsterdam: Studiecentrum Arbeid en Gezondheid, Universiteit van Amsterdam.

Meijman, T.F. & O'Hanlon, J.F. (1984). Workload. An introduction to psychological theories and measurement methods. In P.J.D. Drenth, H. Thierry, P.J. Willems & C.J. de Wolff (Eds.), *Handbook of Work and Organizational Psychology*. (pp. 257-288). New York: Wiley.

Meijman, T.F. & Mulder, G. (1992). Arbeidspsychologische aspecten van werkbelasting. (Workpsychological aspects of work-demand) In P.J.D. Drenth, H. Thierry & Ch.J. de Wolf (Red.), *Nieuw handboek A&O psychologie*, Hoofstuk 2.11. Deventer: Van Loghum Slaterus.

Meister, D. (1976). *Behavioral foundations of system development*. New York: Wiley.

Michaels, R.M. (1962). The effect of expressway design on driver tension responses. *Public Roads, 32*, 107-112.

Michon, J.A. (1971). *Psychonomie Onderweg*. Inaugural lecture, University of Groningen. Groningen, Wolters Noordhoff.

Michon, J.A. (1985). A critical view of driver behavior models: what do we know, what should we do? In L. Evans & R.C. Schwing (Eds.), *Human behavior & traffic safety* (pp. 485-524). New York: Plenum Press.

Michon, J.A. (Ed.) (1993). *Generic Intelligent Driver Support System*. London: Taylor & Francis.

Muckler, F.A. & Seven, S.A. (1992). Selecting performance measures: 'objective' versus 'subjective' measurement. *Human Factors, 34*, 441-455.

Mulder, G. (1980). *The heart of mental effort*. PhD Thesis. Groningen: University of Groningen.

Mulder, G. (1986). The concept and measurement of mental effort. In G.R.J. Hockey, A.W.K. Gaillard and M.G.H. Coles (Eds.), *Energetics and human information processing* (pp. 175-198). Dordrecht, The Netherlands: Martinus Nijhoff Publishers.

Mulder, G. (1995). *The search for energetical indices of mental task load*. Contribution to the EPOS workshop 'Psychophysiology of Mental Resources', September 5-8, Amsterdam.

Mulder, G. & Mulder, L.J.M. (1980). Coping with mental workload. In S. Levine & H. Ursin (Eds.) *Coping and Health* (pp. 233-258). New York: Plenum Press.

Mulder, G. & Mulder, L.J.M. (1981). Information processing and cardiovascular control. *Psychophysiology, 18*, 392-402.

Mulder, L.J.M. (1988). *Assessment of cardiovascular reactivity by means of spectral analysis*. PhD Thesis. Groningen: University of Groningen.

Mulder, L.J.M. (1992). Measurement and analysis methods of heart rate and respiration for use in applied environments. *Biological Psychology, 34*, 205-236.

Mulder, L.J.M., Schweizer, D. & Van Roon, A.M. (1990). *An environment for data reduction, correction and analysis of cardiovascular signals*. Abstract 4th 'Computers in Psychology' workshop, Tilburg, The Netherlands.

Mulders, H.P.G., Meijman, T.F., O'Hanlon, J.F. & Mulder, G. (1982). Differential psychophysiological reactivity of city bus drivers. *Ergonomics, 25*, 1003-1011.

Mulders, H., Meijman, T., Mulder, B., Kompier, M., Broersen, S., Westerink, B. & O'Hanlon, J. (1988). Occupational stress in city bus drivers. In J.A. Rothengatter & R.A. de Bruin (Eds.), *Road user behaviour: theory and research* (pp. 348-358). Assen, The Netherlands: Van Gorcum.

Müller, A., Schandry, R., Montoya, P. & Gsellhofer, B. (1992). Differential effects of two stressors on heart rate, respiratory sinus arrhythmia, and T-wave amplitude. *Journal of Psychophysiology, 6*, 252-259.

Myrtek, M., Deutschmann-Janicke, E., Strohmaier, H., Zimmermann, W., Lawerenz, S., Brügner, G. & Müller, W. (1994). Physical, mental, emotional, and subjective workload components in train drivers. *Ergonomics, 37*, 1195-1203.

Norman, D.A. & Bobrow, D.G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology, 7*, 44-64.

Nygren, T.E. (1991). Psychometric properties of subjective workload measurement techniques: implications for their use in the assessment of perceived mental workload, *Human Factors, 33*, 17-33.

O'Donnell, R.D. & Eggemeier, F.T. (1986). Workload assessment methodology. In K.R. Boff, L. Kaufman & J.P. Thomas (Eds.), *Handbook of perception and human performance. Volume II, cognitive processes and performance*. (pp 42/1-42/49). New York: Wiley.

O'Hanlon, J.F. (1981). Boredom: practical consequences and a theory. *Acta Psychologica, 49*, 53-82.

O'Hanlon, J.F. (1984). Driving performance under the influence of drugs: rationale for, and application of, a new test. *British Journal of Clinical Pharmacology, 18*, 121S-129S.

O'Hanlon, J.F., Haak, T.W., Blaauw, G.J. & Riemersma, J.B.J. (1982). Diazepam impairs lateral position control in highway driving. *Science, 217*, 79-80.

Paas, F.G.W.C., Van Merriënboer, J.G.J. & Adam, J.J. (1994). Measurement of cognitive load in instructional research. *Perceptual and Motor Skills, 79*, 419-430.

Papillo, J.F. & Shapiro, D. (1990). The cardiovascular system. In J.T. Cacioppo & L.G. Tassinary. *Principles of psychophysiology* (pp. 456-512). Cambridge: Cambridge University Press.

Parkes, A.M. (1991). Data capture techniques for RTI usability evaluation. In Commission of the European Communities, *Advanced telematics in road transport*, Proceedings of the DRIVE conference. (pp. 1440-1456). Amsterdam: Elsevier.

Pohlmann, S. & Traenkle, U. (1994). Orientation in road traffic. Age-related differences using an in-vehicle navigation system and a conventional map. *Accident Analysis and Prevention, 26*, 689-702.

Porges, S.W. (1992). Vagal tone: a physiologic marker of stress vulnerability. *Pediatrics, 90*, 498-504.

Porges, S.W., Bohrer, R.E., Cheung, M.N., Drasgow, F., McCabe, P.H. & Keren, G. (1980). New time-series statistic for detecting rhytmic co-occurrence in the frequency domain: the weighted coherence and its application to psychophysiological research. *Psychological Bulletin, 88*, 580-587.

Porges, S.W. & Byrne, E.A. (1992). Research methods for measurement of heart rate and respiration. *Biological Psychology, 34*, 93-130.

Posner, M.I. (1978). *Chronometric explorations of mind*. Hillsdale: Erlbaum.

Pribram, K.H. & McGuiness, D. (1975). Arousal, activation and effort in the control of attention. *Psychological review, 82*, 116-149.

Raggatt, P.T.F. & Morrissey, S.A. (submitted). Stress and fatigue in long-distance coach drivers: cumulative effects of 12 Hr workdays.

Reid, G.B., Shingledecker, C.A. & Eggemeier, F.T. (1981). Application of conjoint measurement to workload scale development. In *Proceedings of the Human Factors Society 25th annual meeting* (pp. 522-526). Santa Monica, CA: Human Factors Society.

Reid, G.B. & Colle, H.A. (1988). Critical SWAT values for predicting operator overload. In *Proceedings of the Human Factors Society 32nd annual meeting* (pp. 1414-1418). Santa Monica, CA: Human Factors Society.

Riedel, W.J. (1991). Eye-movements, expert ratings, weaving and time-to-line-crossing as measures of driving performance and driving performance impairment. In A.G. Gale, I.D. Brown, C.M. Haslegrave, I. Moorhead & S. Taylor (Eds.), *Vision in vehicles -III* (pp. 299-306). Amsterdam: North-Holland.

Rockwell, T.H. (1988). Spare visual capacity in driving-revisited. New empirical results for an old idea. In A.G. Gale, M.H. Freeman, C.M. Haslegrave, P. Smith & S.P. Taylor (Eds.), *Vision in vehicles-II* (pp. 317-324). Amsterdam: North-Holland.

Roscoe, A.H. (1992). Assessing pilot workload. Why measure heart rate, HRV and respiration? *Biological Psychology, 34*, 259-287.

Roscoe, A.H. (1993). Heart rate as psychophysiological measure for in-flight workload assessment. *Ergonomics, 36*, 1055-1062.

Rothengatter, J.A. (1991). Automatic policing and information systems for increasing traffic law compliance. *Journal of applied behavior analysis, 24*, 85-87.

Rouse, W.B., Edwards, S.L. & Hammer, J.M. (1993). Modelling the dynamics of mental workload and human performance in complex systems. *IEEE transactions on systems, man, and cybernetics, 23*, 1662-1671.

Rueb, J., Vidulich, M., Hassoun, J. (1992). Establishing workload acceptability: an evaluation of a proposed KC-135 cockpit redesign. In *Proceedings of the Human Factors Society 36th annual meeting* (pp. 17-21). Santa Monica, CA: Human Factors Society.

Sanders, A.F. (1970). Some aspects of the selective process in the functional visual field. *Ergonomics, 13*, 101-117.

Sanders, A.F. (1983). Towards a model of stress and human performance. *Acta Psychologica, 53*, 61-97.

Schneider, W. & Shiffrin, R.M. (1977). Controlled and automatic human information processing: I. Detection, search and attention. *Psychological Review, 84*, 1-66.

Schneider, W. & Fisk, A.D. (1983). Attention theory and mechanisms for skilled performance. In R. Magill (Ed.), *Memory and control of action*. New York: North-Holland.

Schneider, W., Dumais, S.T. & Shiffrin, R.M. (1984). Automatic and control processing and attention. In R. Parasuraman and D.R. Davies (Eds.), *Varieties of attention*. (pp. 1-27). London: Academic Press.

Settels, J.J. & Wesseling, K.H. (1985). FIN.A.PRES: non-invasive finger arterial pressure waveform registration. In J.F. Orlebeke, G. Mulder & L.P.J. van Doornen (Eds.), *The psychophysiology of cardiovascular control* (pp. 267-283). New York: Plenum Press.

Shiffrin, R.M. & Schneider, W. (1977). Controlled and automatic information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological review, 84*, 127-190.

Sirevaag, E., Kramer, A.F., De Jong, R. & Mecklinger, A. (1988). A psychophysiological analysis of multi-task processing demands. *Psychophysiology, 25*, 482.

Sirevaag, E.J., Kramer, A.F., Wickens, C.D., Reisweber, M., Strayer, D.L. & Grenell, J.F. (1993). Assessment of pilot performance and mental workload in rotary wing aircraft. *Ergonomics, 36*, 1121-1140.

Smiley, A. & Brookhuis, K.A. (1987). Alcohol, drugs and traffic safety. In J.A. Rothengatter & R.A. de Bruin (Eds.), *Road users and traffic safety*. (pp. 83-105). Assen, The Netherlands: Van Gorcum.

Snijders, T.A.B. (1995). *personal communication*.

Sokolov, E.N. (1963). *Perception and the conditioned reflex*. Oxford: Pergamon.

Stein, A.C., Parseghian, Z. & Allen, R.W. (1987). *A simulator study of the safety implications of cellular mobile phone use*. In American Association for Automotive Medicine. (pp. 181-200). Proceedings of the 31st annual conference. Desplaines, IL: AAAM.

Steptoe, A. & Sawada, Y. (1989). Assessment of baroreceptor reflex function during mental stress and relaxation. *Psychophysiology, 26*, 140-147.

Stern, J.A., Boyer, D. & Schroeder, D. (1994). Blink rate: a possible measure of fatigue. *Human Factors, 36*, 285-297.

Steyvers, F.J.J.M. (1993). The measurement of road environment appreciation with a multi-scale construct list. In A.G. Gale, I.D. Brown, C.M. Haslegrave, H.W. Kruysse & S.P. Taylor (Eds.), *Vision in Vehicles-IV*. (pp. 203-212). Amsterdam: North-Holland.

Steyvers, F.J.J.M., Dekker, K., Brookhuis, K.A. & Jackson, A.E. (1994). The experience of road environments under two lighting and traffic conditions: application of a Road Environment Construct List. *Applied Cognitive Psychology, 8*, 497-511.

Teigen, K.H. (1994). Yerkes-Dodson: a law for all seasons. *Theory & Psychology, 4*, 525-547.

Thomas, D.B., Herberg, K.-W., Brookhuis, K.A., Muzet, A.G., Poilvert, C., Tarriere, C., Norin, F., Wyon, D.P., Schievers, G. & Mutschler, H. (1989). *Demonstration experiments concerning driver status monitoring* (Report V1004/DREAM). Köln, FRG, Technische Überwachungs-Verein Rheinland e.V.

Thorsvall, L. & Åkerstedt, T. (1987). Sleepiness on the job: continuously measured EEG changes in train drivers. *Electroencephalography and Clinical Neurophysiology, 66*, 502-511.

Unema, P. (1995). *Eye movements and mental effort*. PhD Thesis, TU Berlin. Aachen, Germany: Verlag Schalter.

Van Boxtel, A. & Jessurun, M. (1993). Amplitude and bilateral coherency of facial and jaw-elevator EMG activity as an index of effort during a two-choice serial reaction task. *Psychophysiology, 30*, 589-604.

Van der Beek, A.J., Meijman, T.F., Frings-Dresen, M.H.W., Kuiper, J.I. & Kuiper, S. (1995). Lorry drivers' work stress evaluated by catecholamines excreted in urine. *Occupational and Environmental Medicine, 52*, 464-469.

Van Ouwerkerk, R., Meijman, Th.F. & Mulder, G. (1994a). Mentale belasting (Mental Workload). In P.C. Buijs, A. van Oosterom & H. Wolvetang (Red.), *Handboek Bedrijfsgezondheidszorg* (Hoofstuk B3-1). Utrecht, The Netherlands: Bunge.

Van Ouwerkerk, R.J., Meijman, T.F. & Mulder, G. (1994b). *Arbeidspsychologische taakanalyse. Het onderzoek van cognitieve en emotionele aspecten van arbeidstaken. (Workpsychological task analysis. Research on cognitive and emotional aspects of tasks of labour)*. Utrecht, The Netherlands: Lemma.

Van Roon, A.M. (in preparation). *Simulation of short-term cardiovascular effects of mental workload*. PhD Thesis, University of Groningen.

Van Winsum, W., Van Knippenberg, C. & Brookhuis, K. (1989). Effect of navigation support on drivers' mental workload. In *Current issues in European transport, Vol I. Guided transport in 2040 in Europe* (pp. 69-84). London: PTRC Education and Research Services.

Van Wolffelaar, P.C., Brouwer, W.B. & Rothengatter, J.A. (1990). *Divided attention in RTI-tasks for elderly drivers* (Report 1006/DRIVAGE/Deliverable TRC1). Haren, The Netherlands: Traffic Research Centre, University of Groningen.

Vaughan, G., May, A., Ross, T. & Fenton, P. (1994). A human-factors investigation of an RDS-TMC system. In ERTICO (Ed.) *Towards an intelligent transport system. Proceedings of the first world congress on applications of transport telematics and intelligent vehicle-highway systems* (pp. 1685-1692). Boston: Artech House.

Veltman, J.A. & Gaillard, A.W.K. (1993). Indices of mental workload in a complex task environment. *Neuropsychobiology, 28*, 72-75.

Veltman, J.A. & Gaillard, A.W.K. (in press). Measurement of pilot workload with subjective and physiological techniques. Paper presented at the annual meeting of the Europe Chapter of the Human Factors and Ergonomics Society, November 1993, Soesterberg, The Netherlands.

Verwey, W.B. (1990). *Adaptable driver-car interfacing and mental workload: a review of the literature* (Report V1041/GIDS/DIA/1). Haren, The Netherlands: Traffic Research Centre, University of Groningen.

Verwey, W.B. (1993a). How can we prevent overload of the driver? In A.M. Parkes & S. Franzén, *Driving future vehicles*. (pp 235-244). London: Taylor & Francis.

Verwey, W.B. (1993b). *Driver workload as a function of road situation, age, traffic density, and route familiarity* (Report IZF 1993 C-11). Soesterberg, The Netherlands: TNO Institute for Perception.

Verwey, W.B. & Veltman, J.A. (1995). *Measuring workload peaks while driving. A comparison of nine common workload assessment techniques* (Report TNO-TM 1995 B-4). Soesterberg, The Netherlands: TNO Human Factors Research Institute.

Vicente, K.J., Thornton, D.C. & Moray, N. (1987). Spectral analysis of sinus arrhythmia: a measure of mental effort. *Human Factors, 29*, 171-182.

Vidulich, M.A. & Tsang, P.S. (1986). Techniques of subjective workload assessment: a comparison of SWAT and the NASA-bipolar methods. *Ergonomics, 29*, 1385-1398.

Vidulich, M.A. & Wickens, C.D. (1986). Causes of dissociation between subjective workload measures and performance. Caveats for the use of subjective assessments. *Applied Ergonomics, 17*, 291-296.

Vivoli, G., Bergomi, M., Rovesti, S., Carrozzi, G. & Vezzosi, A. (1993). Biochemical and haemodynamic indicators of stress in truck drivers. *Ergonomics, 36*, 1089-1097.

Volkerts, E.R., Louwerens, J.W., Gloerich, A.B.M., Brookhuis, K.A. & O'Hanlon, J.F. (1984). *Zopiclone's residual effect upon actual driving performance versus those of Nitrazepam and Flunitrazepam* (Report VK 84-10). Haren, The Netherlands: Traffic Research Centre, University of Groningen.

Volkerts, E.R., Brookhuis, K.A. & O'Hanlon, J.F. (1987). *Comparison of the effects of Buspirone 5 mg and 10 mg, Diazepam 5 mg and Lorazepam 1 mg (t.i.d.) upon actual driving performance* (Report VK 87-02). Haren, The Netherlands: Traffic Research Centre, University of Groningen.

Waterink, W. & Van Boxtel, A. (1994). Facial and jaw-elevator EMG activity in relation to changes in performance level during a

sustained information processing task. *Biological Psychology, 37*, 183-198.
Wickens, C.D. (1984). Processing resources in attention. In R. Parasuraman and D.R. Davies (Eds.). *Varieties of attention*. (pp. 63-102). London: Academic Press.
Wickens, C.D. (1991). Processing resources and attention. In D.L. Damos (Ed.), *Multiple-task performance*. (pp. 3-34). London: Taylor & Francis.
Wickens, C.D. (1992). *Engineering psychology and human performance*. New York: HarperCollins.
Wiener, E.L. (1987). Application of vigilance research: rare, medium, or well done? *Human Factors, 29*, 725-736.
Wientjes, C.J.E. (1992). Respiration in psychophysiology: methods and applications. *Biological Psychology, 34*, 179-203.
Wientjes, C.J.E. (1993). *Psychological influences upon breathing: situational and dispositional aspects*. PhD thesis, Soesterberg, The Netherlands: TNO Institute for Perception.
Wierwille, W.W. & Casali, J.G. (1983). A validated rating scale for global mental workload measurement application. In *Proceedings of the Human Factors Society 27th Annual Meeting* (pp. 129-133). Santa Monica, CA: Human Factors Society.
Wierwille, W.W., Rahimi, M. & Casali, J.G. (1985). Evaluation of 16 measures of mental workload using a simulated flight task emphasizing mediational activity. *Human Factors, 27*, 489-502.
Wierwille, W.W. & Eggemeier, F.T. (1993). Recommendation for mental workload measurement in a test and evaluation environment. *Human Factors, 35*, 263-281.
Wijers, A.A. (1989). *Selective visual attention: an electrophysiological approach*. PhD thesis, University of Groningen.
Wildervanck, C., Mulder, G. & Michon, J.A. (1978). Mapping mental load in car driving. *Ergonomics, 21*, 225-229.
Wilhelm, B. & Wilhelm, H. (1995). Das Pupilverhalten verrät Übermüdung. *Zeitschrift für Verkehrssicherheit, 41*, 116-118.
Wilson, G.F. (1992). Applied use of cardiac and respiration measures: practical considerations and precautions. *Biological Psychology, 34*, 163-178.
Wilson, G.F. & Eggemeier, F.T. (1991). Psychophysiological assessment of workload in multi-task environments. In D.L. Damos (Ed.), *Multiple-task performance* (pp. 329-360). London: Taylor & Francis.
Yeh, Y.Y. & Wickens, C.D. (1988). Dissociation of performance and subjective measures of workload. *Human Factors, 30*, 111-120.
Yerkes, R.M. & Dodson, J.D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology, 18*, 459-482.

Zeier, H. (1979). Concurrent physiological activity of driver and passenger when driving with and without automatic transmission in heavy city traffic. *Ergonomics, 22*, 799-810.

Zijlstra, F.R.H. (1993). *Efficiency in work behavior. A design approach for modern tools*. PhD thesis, Delft University of Technology. Delft, The Netherlands: Delft University Press.

Zijlstra, F.R.H. & Van Doorn, L. (1985). *The construction of a scale to measure perceived effort*. Delft, The Netherlands: Department of Philosophy and Social Sciences, Delft University of Technology.

Zijlstra, F. & Meijman, T. (1989). Het meten van mentale inspanning met behulp van een subjectieve methode (measurement of mental effort with a subjective method). In T. Meijman (Ed.), *Mentale belasting en werkstress. Een arbeidspsychologische benadering*. (pp. 42-61). Assen, The Netherlands: Van Gorcum.

Zijlstra, F. & Mulder, G. (1989). Mentale belasting: theoretische gezichtspunten en overzicht van meetmethoden (mental workload: theoretical points-of-view and an overview of measurement methods). In T. Meijman (Ed.), *Mentale belasting en werkstress. Een arbeidspsychologische benadering*. (pp. 21-41). Assen, The Netherlands: Van Gorcum.

8
Appendices

Appendix 1: De Waard, D. (1991). Driving behaviour on a high-accident-rate motorway in the Netherlands. In C. Weikert, K.A. Brookhuis and S. Ovinius (Eds.), *Man in complex systems*, Proceedings of the Europe Chapter of the Human Factors Society Annual Meeting. Work Science Bulletin 7 (pp 113-123). Lund, Sweden: Work Science Division, Department of Psychology, Lund University.

Appendix 2: De Waard, D., Jessurun, M., Steyvers, F.J.J.M., Raggatt, P.T.F., Brookhuis, K.A. (1995). Effect of road layout and road environment on driving performance, drivers' physiology and road appreciation. *Ergonomics*, *38*, 1395-1407. Published by Taylor & Francis Ltd.

Appendix 3: Brookhuis, K.A., De Vries, G. & De Waard, D. (1991). The effects of mobile telephoning on driving performance. *Accident Analysis and Prevention, 23*, 309-316. Published by Pergamon, Elsevier Science Ltd.

Appendix 4: De Waard, D., Van der Hulst, M. & Brookhuis, K.A. (submitted). Elderly and young drivers' reactions to an in-car enforcement and tutoring system.

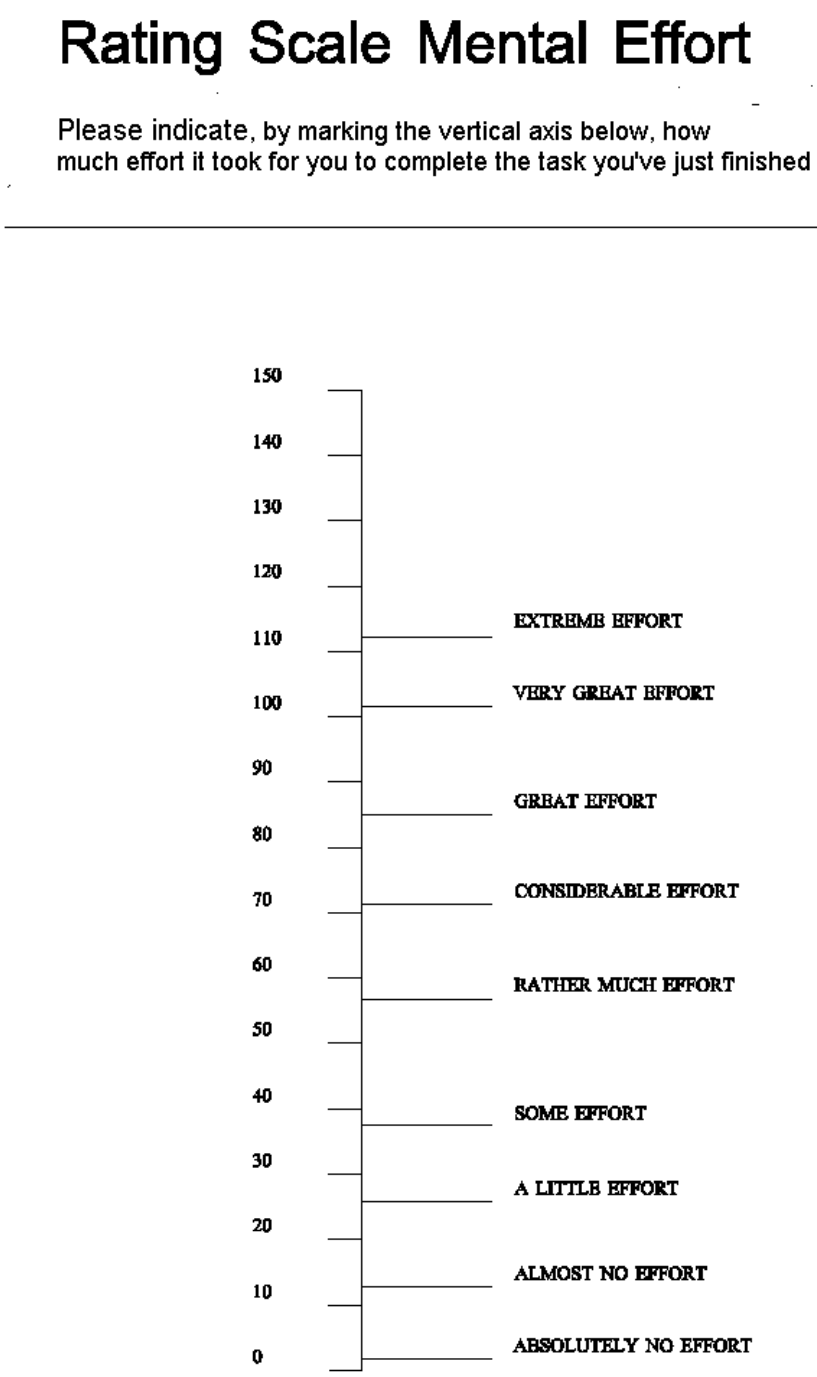
Appendix 5: De Waard, D. & Brookhuis, K.A. (1991a). Assessing driver status: a demonstration experiment on the road. *Accident Analysis and Prevention, 23*, 297-307. Published by Pergamon, Elsevier Science Ltd.

Appendix A: RSME, effort rating scale.

Appendix B: Bartenwerfer's activation scale.

Appendix A

RSME Rating Scale Mental Effort



Rating Scale Mental Effort (Zijlstra, 1993).
Score is indicated by the digits on the left, the official scale is sized such that 150 equals 150 cm from origin to top (cm = centimetres, 150 cm = 0.15 metres)

to chapter 7 (references)
back to thesis summary

I like to hear from you, so if you find this information useful, a short message is very much appreciated. For more information you can also contact me.

De Waard, D. (1996). *The measurement of drivers' mental workload*. PhD thesis, University of Groningen. Haren, The Netherlands: University of Groningen, Traffic Research Centre.

 Back to my HomePage