Assessment of Subjective Workload in Traffic Situations using Video Fragments

Graduation Committee:

Name	Function	Company
Prof. dr. ir. B. van Arem	Chairman	TU Delft
Dr. R.G. Hoogendoorn	First daily supervisor	TU Delft
Dr. Ir. C.N. van Nes	Second daily supervisor	SWOV
M.W.T. Christoph, Msc.	Third daily supervisor	SWOV
Prof. dr. K.A. Brookhuis	External supervisor	TU Delft
Ir. P.B.L. Wiggenraad	Graduation coordinator	TU Delft

R.G.C. Hagendoorn
1371487
Msc. Civil Engineering

May, 2014

TU Delft

SWOV
WETENSCHAPPELIJK ONDERZOEK VERKEERSVEILIGHEID

Preface

This report was made as a fulfilment for obtaining the MSc degree in Civil Engineering. In this report the findings and results of the activities during the graduation period are described. This research was made possible through cooperation with the Stichting Wetenschappelijk Onderzoek Verkeersveiligheid (SWOV), whom financed this project and allowed the use of their naturalistic driving-dataset.

This preface is followed by an extensive summary in which the content and structure of the report is reproduced in a more concise form and could be read as an alternative to the main text without missing important parts of the content. This summary is created out of the individual chapter summaries which can be found at the end of each chapter.

Abstract

Mental workload is an important subject in traffic safety research, however workload measurements are expensive and time intensive. To this end a measurement method is developed which uses short video fragments of driving situations to the subject, after which they are asked about their perceived mental effort if they would be driving through those situations themselves. In this way, 32 video fragments taken from the naturalistic driving dataset are shown to 60 participants, providing information on several traffic variables, which have an effect on workload. The method managed to find significant effects for the effect of traffic density, age, weather and presence of Heavy Goods Vehicles and Vulnerable Road Users on mental workload, which can also be found in workload studies performed in the past. The method developed in this report appears to be promising as a sensitive and valid method for determining workload. However, in order to adequately attain its effectiveness, more research is needed.

Keywords

Mental workload, Subjective workload, Video Fragments, Naturalistic Driving, Rating Scale Mental Effort (RSME)

Executive Summary

Driving involves high fluctuations in mental workload, since a wide variety of demands is placed on the driver in a high sequence. At a high workload it is possible for the driver to become overloaded and unable to respond to new information. Low workload can result in boredom, decreased situational awareness and an overall reduction in alertness. Both high and low workload are causes of driver's inattention, which is seen as one of the primary causes of traffic accidents. In order to better prevent traffic accidents, it is important to understand the cause and mechanics behind mental workload. For this reason, workload is the subject of a large body of research. Mental workload research is generally performed using either instrumented vehicles or driving simulators. These studies are generally time-intensive and expensive to carry out. This research describes the development of a workload measure method which has the primary benefit that it is very easy and inexpensive to implement.

The objective of this research is to develop a method with which mental workload can be measured using video fragments, and to gauge its usability in mental workload research though the use of an experimental application. In order to work towards this objective a number of research questions are created:

1. What is mental workload and how can it be measured?
2. Which aspects of car driving can result in increased mental workload?
3. How can video fragments be used to measure mental workload?
4. What is the validity of the measurement method?
5. How does the measurement method compare to other workload measurement methods?

In this study a new method for measuring mental workload is developed. In the method, proposed in this research, short video fragments are shown to the participants. In these video fragments traffic situations are shown from the viewpoint of a car driver. After watching each video fragment the participant rates their perceived mental effort on a subjective rating scale, imagining that they are driving the vehicle in that particular situation themselves. These images were taken from a naturalistic driving study, performed at the Stichting Wetenschappelijk Onderzoek Verkeersveiligheid (SWOV), in which participants are provided with a vehicle, instrumented with several sensors and cameras for a period of 4 weeks. This vehicle registers driving attributes such as speed and acceleration continuously, and captures video images of the driver and front view through the vehicles windshield. This created hundreds of hours of footage, of which a small selection is made and used in this research.

Mental workload and measurement

While there is no agreed upon definition of workload which is used by all researchers, there are some common elements. Overall, mental workload is described as the portion of the information processing capacity that is employed to achieve task performance. Two factors are important in the determination of workload: the demand of the task and the capability of the person performing the task. Task demand is a processing requirement to perform a task with desired performance, independent of the capability of the individual performing the task. Task demand is determined for a large part by the complexity of the task, which is reflected in the number of processing stages that are required to perform the task. Task difficulty is the processing effort, specific for an individual,

which is required for the task and is determined by factors such as task demand, processing capacity, experience and state of mind.

O'Donnel and Eggemeier (1986) and De Waard (1996) mentioned a number of measurement evaluation criteria which are used in order to be able to judge measurement methods on their applicability. These criteria could eventually be used to evaluate on the measurement method developed in this research. Sensitivity is determined by the extent to which the measurement method can detect changes in the level of mental workload at different levels of task performance. Diagnosticity relates to ability of the measurement method to discriminate the workload from different resource pools, which are the perceptual, central processing and motor input (Wickens, 1991). Primary task intrusion refers to the degree to which the measurement method intrudes upon the primary task, which constitutes to safety operating the vehicle in most driving workload researches. Implementation requirements are the expertise, time and financial requirements that are necessary to apply the measurement method. Operator acceptance refers to the willingness of the subject to participate in the experiment and to what degree they think the research is valid. Selectivity is the degree to which the measurement method is sensitive to specifically the trait that is being researched. Finally, bandwidth and reliability refer to the consistency and applicability of the method over different repeated measures and applications at different performance levels.

Four different categories of measurement methods are discussed: subjective measures, primary- and secondary task performance measures, and physiological measures. When using subjective measures, the participant is asked to rate their perceived effort after completing the task. With uni-dimensional subjective measures a single unit for the perceived effort of the participant is obtained, while multi-dimensional subjective measures have different categories for physical workload, cognitive workload, frustration level and performance level. Subjective measures generally have a good sensitivity and low primary task intrusion, but can be subject to bias. Primary task performance measures use measures in task performance such as lane keeping or speed maintaining to determine workload. These measures are generally insensitive to changes in workload during optimal performance. In secondary task performance an additional task is included, of which common examples are calculation- or reaction tasks. A reduction in secondary task performance could be measured during high workload situations, where more attention needs to be directed towards the primary task. For this, it is important that the secondary task does not intrude upon the primary task, so only the secondary and not the primary task performance degrades during these situations. Physiological measures are observations of the operator's state, through measures of physiological processes. Commonly occurring examples are the use of skeletal muscle or heart monitoring devices. It is important that a good control over the physical activity of the subject is maintained, since these measures are often sensitive to this. Table 1 displays the different measurement method categories and their strengths and weaknesses. Quality of the different methods in the table are indicated by a +, – or 0. A + indicates a positive index, 0 an average index and – a negative index. Indexation is done on a relative level, e.g. subjective and secondary task performance score 0 at sensitivity although they are considered generally sensitive, since physiological measures have such high sensitivity.

Table 1 - Measurement Method Evaluations

	Sensitivity	Diagnosticity	Prim. task intrusion	Implement. req.	Operator acceptance	Selectivity	Bandw. & reliability
Subjective	0	- / 0*	+	+	+	+	0
Prim. task	-	-	+	+	+	+	0
Sec. Task	0	+	-	0	0	0	+
Physiological	+	+	0	-	0	-	+

**- for uni-dimensional scales and 0 for multi-dimensional scales*

Variables in traffic affecting mental workload

The evaluation of the measurement method developed in this research is done by comparing its findings to effects found in mental workload studies performed in the past. For this, it is important to know what aspects of traffic have an effect on workload, in order to know which factors to include in the experiment. To do this, a review of the literature on the variables present in traffic which could act as a predictor for mental workload is made. The review discusses the effects of driver characteristics, road geometry, environmental aspects and vehicle characteristics. Among driver characteristics the driver's age, driving experience, gender, personality and familiarity with the surroundings were studied. Studies finding a relation between the driver's age and driving experience with mental workload were found. For the driver's gender the results were mixed, some studies reported on significant effects, however no consensus was found. In a previous study significant effects were found for unfamiliarity with the surroundings, but only when the subject is not using a route-guidance system.

Three aspects of road geometry were studied: lane width, road curvature and whether the road is a single- or dual carriageway road. In previous study it was found that traffic on the opposing lane results in an increased steering demand away from the center of the road, resulting in higher workload. Having a decreased road width also results in a higher steering demand, also resulting in higher workload. Size of the curvature and speed over the curve is also resulting in a high workload. Furthermore it was found that driving over an uncontrolled intersection and roundabout resulted in a high workload.

For environmental aspects, a number of aspects which may occur in the driving environment were studied. Among this are the presence of Vulnerable Road Users (VRUs) and Heavy Goods Vehicles (HGVs), effects of changes in lighting and weather, environment complexity, and behaviour and number of other vehicles on the road. While no previous research studying the effect of the presence of VRUs was found, a number of studies assumed their presence results in an increase in workload. For the presence of HGVs a couple of studies showed their effect on workload. A reduction in lighting and poor weather results in reduced vision from the driver, which results in higher workload. Traffic density has been found to be an important aspect of traffic in a number of studies; not only the density of the shared lanes, also the density of the opposing lanes was determined to have an effect of workload. In this report, environmental complexity is determined as all the aspect of the traffic surrounding that could have an effect on workload but do not directly interfere with the driving task. Examples are tall buildings, presence of roadside advertisement billboards and accidents happening on the other driving lane. Increased environmental complexity is a reason why urban traffic correlates to an overall higher workload requirement, along with the increase complexity of the vehicle interactions. Higher order driving maneuvers are considered one of the most demanding

aspects of driving. Examples such as merging, overtaking and lane changing are very complex, since the driver needs to make decisions on both longitudinal as well as lateral control in in a short period of time, while maintaining awareness of surrounding vehicles.

The last category which was studied are the vehicle characteristics, in which attention was mostly spend towards advanced driver assistance systems (ADAS). A number of ADAS systems have the potential to reduce driver workload by taking over certain aspects of the driving task, e.g. speed maintaining or navigating. Operating the device while driving can substantially increase workload however, which is especially the case with the use of mobile phones during driving.

Experimental Setup

In the proposed research the participant watches a series of video fragments, each containing a combination of a number of variables of which in the past was found to have an influence on mental workload. For these fragments the participant is asked to place themselves into the car driver's position. After each video fragment the participant indicates a rating on a subjective rating scale.

A total of 60 participants divided over 3 age categories: a young group (18-25 years), a middle aged group (30-50 years) and an elderly group (65+); participate in the experiment. Besides their age, information is asked on their: gender, driving experience in years, yearly kilometrage and the frequency with which they use the motor way. The participants are presented the video fragments through a software application created for the sole purpose of this experiment. This application presents the video fragments (15 seconds in length) one by one to the participant, allowing them to fill in a rating on the scale after each individual fragment. The subjective scale selected for this experiment is the Rating Scale Mental Effort (RSME; Zijlstra, 1993), because of its uni-dimensionality and sensitivity in high performance driving situations (De Waard, 1996). The experiment is mostly self-paced, but does not allow the participant to fill in a rating until the full 15 seconds of the video are played out, and does not allow the start of the next video until the slider of the RSME is moved. All participants rate all video fragments, which are presented to the participants in a random order as a means of counterbalancing. Figure 1 shows the interface of the program used in the experiment.

Figure 1 - Experiment Software

The experiment features two designs. The primary design is a factorial design of the three variables: traffic density, number of lanes and the presence of HGVs. For these fragments only motor way situations are selected and each combination is included twice. The reason to only include motor way situations is because they can easily expressed as a combination of a limited amount of variables and suffer little from environmental complexity, which is difficult to quantify due to the wide range of possible sources of complexity. The secondary design contains pairs of near-identical situations of which the only difference is the addition of either a single or multiple vulnerable road users (VRUs) or poor weather (rain) to one of the pairs. This allows for the study on the effect of the presence of these variables on mental workload, as well as a secondary purpose which is the masking of the primary design. When the experiment would contain solely slight variations in the motor way situations, the participant may catch on to the purpose of the experiment which could influence their decisions.

Before the participant is allowed to start the experiment, a short introduction is provided to them. This introduction services several purposes: it explains the procedure of the experiment, familiarizes the participant with the use of the RSME scale, and gives a reference for the use of the RSME for the first couple of videos. Participants during the pilot indicated that especially for the first couple of videos they have trouble placing their RSME marker, having yet to build a frame of reference for themselves. After a small number of videos the participants are able to compare the new videos to the previous ones and are better able to secure a rating. This is done by showing the participant three example situations created for the purpose of the introduction; an expectantly low effort situation, a high effort situations and a situation which falls somewhere in between these two. For each of these examples an indication is given on the RSME scale where the situations could be scored. After the introduction is provided the participants start on the actual experiment, which they perform independently.

Results

The results of the experiment were analyzed using a Repeated Measures ANOVA (rANOVA). Significant main effects ($p<0.05$) in the motor way situations were found for: traffic density, presence of HGVs and Age of the participant. Participants found that the high density situations required much more effort than the low density situations. The presence of HGVS was also found to have a statistical significant influence, however the number of lanes did not result in significant results. The main effects of the within subject factors are shown in figure 2.

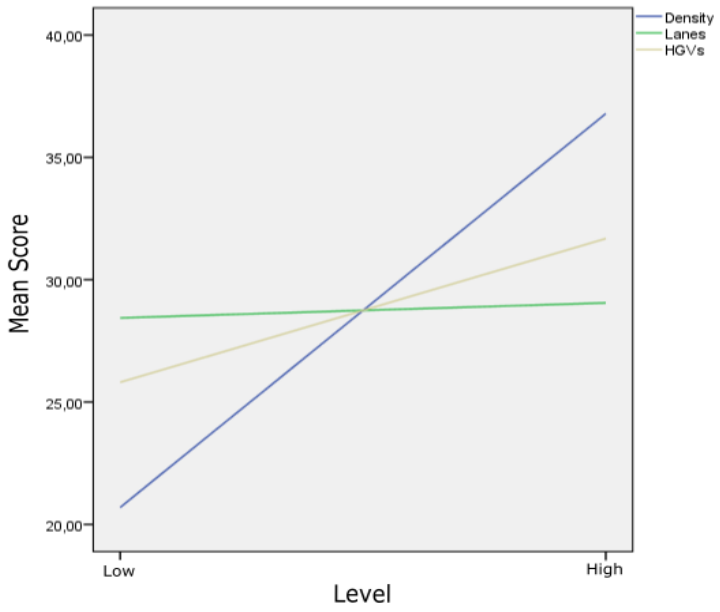


Figure 2 - Main effect within-subject variables

In the experiment the young group indicated to have the highest mean workloads, followed by the elderly group, and the middle-aged group indicated to require the least amount of effort in the explored situations. This effect is explainable; younger drivers lack the experiences and automatisms present in the other groups, while older drivers have a reduced processing capacity as a result of the aging process. However in studies using subjective workload measures it is often prevalent that the younger group overestimates their driving skill, which was not found in this research. The main effect of age is shown in figure 3.

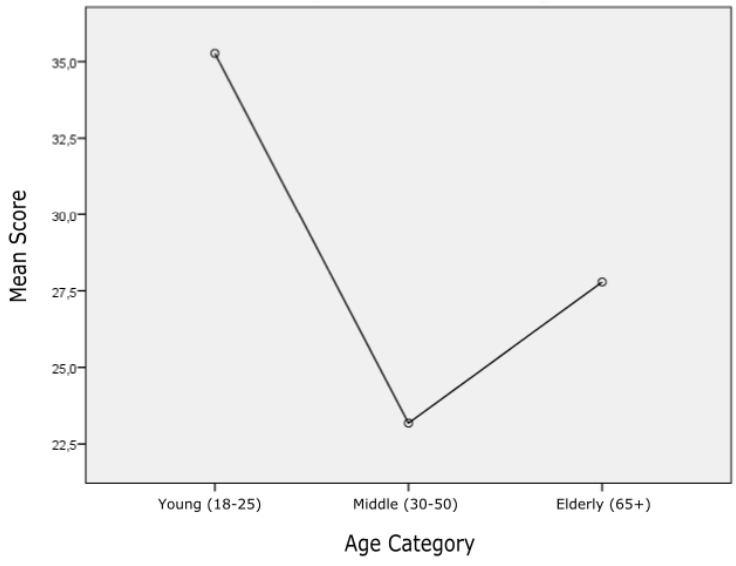


Figure 3 - Main effect age

Interaction effects were found for Density x HGVs, Age x Lanes and Age x Density. The effect of HGVs was greatest in low density situations. The younger group found 3 lane situations to be more effortful than 2 lane situations at high density, while the other two groups found the 2 lane situations to be more effortful. Furthermore the younger group showed a higher relative increase in mean

workload as a result of increased density than the other two groups did. An overview of all results is shown in table 2.

Table 2 - Results r-ANOVA (values are standardized effect sizes (η^2))

	Main Effect	Density	Lanes	HGVs	Age
Density	0.749**	x	0.001	0.162**	0.121*
Lanes	0.014	0.001	x	0.035	0.142*
HGVs	0.465**	0.162**	0.035	x	0.065
Age	0.163*	0.111*	0.142*	0.065	x

In table 2 the standardized effect sizes (partial eta-square) are shown, with an asterisk indicating a significance level of p=0.05 or lower and a double asterisk indicating p=0.01 or lower. According to Kirk (1996) an effect size is small when it is has a value of 0.01, medium at 0.06 and large at 0.14.

rANOVA on the secondary videos showed a main effect of the presence of VRUs and poor weather, with an interaction effect found for VRU x Age; again having the younger group indicate a relatively higher increase in mean workload in the more difficult situation. A rANOVA performed including the gender showed an interaction effect between age and gender, but no main effect. No significant effects as a result of difference in kilometrage were found.

Evaluation of the measurement method

Because the method used in this research is a relatively untested way to measure mental workload, attention is paid towards the evaluation of this method. The method is judged on its test validity and the experiment is judged on its experimental validity. There are two reasons which could result in a low degree of test validity. Only a self-report measure is applied, while in workload research generally two or more measurement methods are used. Self-report measures are subject to bias, as people may overestimate their driving ability. Furthermore it is unknown what the effect is of having the subject watch video fragments of traffic situations, instead of having them participate in traffic themselves. These two effects may amplify each other, since a lack of driving skill from the participant is not necessarily correlated to a decrease in the driver's performance when using this method.

An approximation of the test validity is made by comparing the effects which are found in prior studies which employ methods of obtaining workload which contain better validity. Influence of density, HGVs and age on mental workload were found, which is in line with the results obtained through this method. No information was found on the effect of the number of lanes, which neither confirms nor denies the results found here.

For the experimental validity, there are a couple of concerns with internal and external validity. The lack of control over the variables and presence of confounding variables may cause a reduction in internal validity. Because of the nature of naturalistic driving, there is no control over the traffic situations in which the vehicle drives and the variables that occur during these situations. Variable control during the creating of the video fragments was mostly focused on the prevention of the confounding caused by road curvature and maneuvers such as overtaking. Compromises have been made mostly on the presence of roadside distractions, especially in sparingly occurring combinations

of variables. A series of rANOVAs was performed to test the influence of confounding, by instead of using average scores of the two identical situations use different combinations of single video scores. The results show some difference in the statistical significance of interaction effects that were found, however the main effect of density, HGVs and age the results are still significant for each combination.

For the external validity a number of concerns are identified. Information on a number of personal characteristics was missing, such as social economic status or marital status, of which it would be preferred to have an equal distribution on the sample. Furthermore a bias is likely found towards a higher level of education, since a lot of participants were recruited among employees and students at the VU University in Amsterdam. Recruitment was centered around the cities The Hague and Amsterdam, which results in a high prevalence of participants whom are likely familiar driving in city surroundings. The possible bias introduced by this is however not testable without performing a similar study using a sample with different characteristics.

As a result of the use of a repeated measures design, a bias may be created in the results. A number of tests were performed in order to test if any order- or learning effect can be demonstrated. No significant effects were found as a result of differences in the difficulty of the first video, the order and distance between two coupled video fragments or learning effects.

Conclusion

The objective of this research is to develop a method through which mental workload can be measured using video fragments, and to gauge its usability in mental workload research though the use of an experimental application. To this end a number of research questions were created:

1. What is mental workload and how can it be measured?
2. What aspects of car driving can result in increased mental workload?
3. What is the best method to measure mental workload using video fragments?
4. What is the validity of the measurement method?
5. How does the measurement method compare to other workload measurement methods?

The first three questions are answered in detail in the first three paragraphs and the fourth question was answered in the previous paragraph. The fifth question can be answered by taking a look at the measurement evaluation criteria displayed in table 1. The method has shown to be surprisingly sensitive. This is shown from the high number of effects which were found, even though the driving difficulty in all situations was low to moderate at best and driving performance was always (near) optimal. Diagnosticity of the method is nonexistent, since a uni-dimensional scale is used. It is possible for there to be occurring some primary task intrusion, when the primary task is considered to be trying to place yourself into the driver's position. Because of the short duration and quick sequence of the different video fragments, the scale appearing in between every two fragments can result in primary task intrusion. The low cost and time investment required to implement this method, especially when considered with conventional workload measurement methods, are the primary reason to employ this technique. Because of the short time requirement operator acceptance is high, though the participant is not always convinced on the usefulness of the method, as it was often noted that there were no truly difficult situations in the experiment. The selectivity of the method is unknown, for this comparison will have to be made using a method which has been

found to actually measure workload. Since it is the first time this method has been applied for determining workload, no information is known on the bandwidth and reliability of the method.

The method developed in this research shows potential to be used as a valid and sensitive workload measurement method. The method demonstrates its sensitivity by being able to find significant changes in workload as a result of changes in present variables. While no definite validity could be assigned to the method, it is still recommend using this method for workload assessment in naturalistic driving studies. The method distinguishes itself in its ease of implementation, requiring only a short period of time for each participant and no specific equipment or instruments. Subsequent application of the method, especially in combination with other workload measurement methods, will help strengthen its validity.

Executive summary - Nederlands

Autorijden betreft grote fluctuaties in mentale taaklast, aangezien in korte tijd een grote variëteit aan taken wordt ondernomen door de bestuurder. Bij een hoge taaklast is het mogelijk dat de bestuurder overladen wordt en niet in staat is om te reageren op nieuwe informatie. Lage taaklast kan resulteren in verveling, verminderde bewustheid en een vermindering in alertheid. Zowel hoge als lage taaklast zijn oorzaken van onoplettendheid, wat gezien wordt als een van de voornaamste oorzaken van verkeersongelukken. Om beter in staat te zijn om ongelukken te voorkomen is het belangrijk om de oorzaak en mechanismes achter mentale taaklast te begrijpen. Om deze reden is taaklast het onderwerp van een groot aantal onderzoeken. Normaal gesproken wordt onderzoek naar taaklast uitgevoerd met ofwel geïnstrumenteerde voertuigen of rijsimulators. Deze studies zijn tijdsintensief en duur om uit te voeren. In dit onderzoek wordt de ontwikkeling van een taaklast meetmethode ontwikkelend, waarbij het voornaamste voordeel de geringe kosten en tijdsinvestering zijn.

In deze studie wordt een nieuwe methode voor het meten van taaklast ontwikkeld. Met deze methode worden korte video fragmenten weergegeven aan de deelnemer, waarin verkeerssituaties gezien worden vanuit het oogpunt van een autobestuurder. Na het bekijken van een fragment geeft de deelnemer op een subjectieve meetschaal zijn verwachte mentale inspanning aan, wanneer deze zich voorstelt dat hij zelf in deze situatie rijdt. Deze beelden zijn opgenomen in een natuurlijk rijgedrag onderzoek, uitgevoegd door de Stichting Wetenschappelijk Onderzoek Verkeersveiligheid (SWOV), waarin deelnemers gedurende een periode van 4 weken een voertuig wordt uitgeleend uitgerust met verscheiden sensoren en camera's. Dit voertuig registreert rijeigenschappen als snelheid en versnellingen continu, en neemt video beelden op van de bestuurder en vooruitzicht door de voorruit van de auto. Op deze manier wordt honderden uren aan beeldmateriaal gecreëerd, waarvan een selectie is gemaakt en gebruikt is in dit onderzoek.

Het doel van dit onderzoek is het ontwikkelen van een methode waarmee mentale taaklast kan worden gemeten gebruik makend van video fragmenten, en het peilen van de bruikbaarheid van deze methode in mentale taaklast onderzoek door uitvoering van een experiment. Hiertoe zijn een aantal onderzoeksvragen ontwikkeld:

1. Wat is mentale taaklast en hoe kan het gemeten worden?
2. Welke aspecten van autorijden kunnen tot verhoogde mentale taaklast leiden?
3. Hoe kunnen video fragmenten gebruikt worden om taaklast te meten?
4. Wat is de validiteit van de meetmethode?
5. Hoe verhoudt de meetmethode zich vergeleken met andere methoden voor het meten van taaklast?

Mentale taaklast en meetmethoden

Hoewel er geen algemene definitie is die gebruikt wordt door alle onderzoekers, zijn er een aantal gemeenschappelijke kenmerken. Over het algemeen wordt taaklast beschreven het deel van de informatie verwerkingscapaciteit die wordt gebruikt voor taakuitvoering. Twee factoren zijn van belang bij het bepalen van de taaklast: de task demand en de vaardigheid en ervaring van de persoon die de taak uitvoert. Task demand wordt voornamelijk bepaald door de complexiteit van de taak.

Task difficulty is de cognitieve inspanning, specifiek voor een individu, die benodigd is om de taak uit de voeren. En wordt bepaald door factoren als task demand, proces capaciteit, ervaring en gemoedstoestand.

O'Donnel en Eggemeier (1986) en De Waard (1996) noemden een aantal evaluatiecriteria die kunnen worden gebruikt om meetmethoden op hun bruikbaarheid te beoordelen. Deze criteria kunnen uiteindelijk gebruikt worden om een oordeel te maken over de meetmethode die in dit onderzoek wordt ontwikkeld. Sensitivity wordt bepaald door de mate waarin de meetmethode veranderingen in taaklast kan detecteren op verschillende prestatieniveaus. Diagnosticity heeft betrekking op het vermogen van de meetmethode om onderscheidt te maken tussen de verschillende soorten taaklast: perceptueel, informatieverwerking en motoriek (Wickens , 1991). Primary task intrustion heeft betrekking op de mate waarin de meetmethode invloed heeft op de uitvoering van de primaire taak. Implementation requirements zijn de expertise, tijd en financiële benodigdheden voor het toepassen van de meetmethode . Operator Acceptance is de bereidheid van de deelnemer om deel te nemen aan het experiment en in welke mate zij denken dat het onderzoek geldig is. Selectivity is de mate waarin de meetmethode daadwerkelijk gevoelig is voor de eigenschap die wordt onderzocht. Bandwith and reliability verwijzen naar de consistentie en toepasbaarheid van de methode over herhaalde metingen en toepassingen op verschillende prestatieniveaus.

Vier categorieën meetmethoden worden onderscheidt: subjectieve methoden, primaire- en secundaire taak prestatie, en fysiologische meetmethoden. Bij subjectieve meetmethoden wordt de deelnemer na het uitvoeren van de taak gevraagd om zijn inspanningsniveau aan te geven op een meetschaal. Bij uni-dimensionele schalen wordt de inspanning in een enkele eenheid verkregen, terwijl bij multi-dimensionele schalen wordt gevraagd naar verscheidene categorieën als fysieke last, mentale last, frustratie niveau en prestatieniveau. Subjectieve meetmethoden hebben over het algemeen een goede sensitivity en weinig primary task intrustion, maar kunnen beïnvloed worden door persoonlijke bias. Primaire taakprestatie meetmethoden gebruiken eigenschappen als laterale beweging, en snelheidshandhaving om taaklast te bepalen. Deze methoden zijn over het algemeen ongevoelig voor veranderingen in taaklast bij een optimaal prestatieniveau. Bij secundaire taakprestatie meetmethoden wordt een extra taak toegevoegd aan de rijtaak, met als veelvoorkomende voorbeelden reken- of reactietaken. Een vermindering in secundaire taak prestatie kan worden gemeten tijden hoge taaklast situaties, wanneer more aandacht besteedt moet worden richting de primaire taak. Hiervoor is het belangrijk dat de secundaire taak de uitvoering van de primaire taak niet beïnvloed, zodat alleen de secundaire taakprestatie verminderd tijdens deze situaties. Fysiologische meetmethoden zijn observaties van de deelnemers staat, door het meten van fysiologische processen. Veel voorkomende voorbeelden zijn het meten van skeletspieractivatie of hartactiviteit. Het is belangrijk dat goede controle over de fysieke activiteit van de deelnemer wordt bewaakt, omdat deze methoden vaak gevoelig zijn hiervoor. In tabel 1 worden de verschillende methoden en hun voor- en nadelen weergeven. De kwaliteit van de verschillende methoden zijn aangegeven met een +, - of 0. Een +geeft een positieve index aan, een 0 een gemiddelde index en een – een negatieve index. Indexatie is gedaan op een relatief niveau, e.g. subjectieve en secundaire taak prestatie krijgen een score van 0 toegewezen, hoewel deze een goede sensitivity bezitten, omdat deze minder gevoelig zijn dan fysiologische meetmethoden.

Tabel 1 - Meetmethode Evaluaties

	Sensitivity	Diagnosticity	Prim. task intrusion	Implement. req.	Operator acceptance	Selectivity	Bandw. & reliability
Subjectief	0	- / 0*	+	+	+	+	0
Prim. Taak	-	-	+	+	+	+	0
Sec. Taak	0	+	-	0	0	0	+
Fysiologisch	+	+	0	-	0	-	+

**- voor uni-dimensionele schalen en 0 voor multi-dimensionele schalen*

Variabelen in het verkeer die invloed hebben op taaklast

De evaluatie van de meetmethode die ontwikkelt wordt in dit onderzoek wordt gedaan door het vergelijk van de effecten gevonden in deze studie met degenen die in studies in het verleden zijn gevonden. Hiervoor is het belangrijk om te weten welke verkeersaspecten een effect hebben of taaklast. Hiertoe wordt eerst een review gemaakt van de literatuur met betrekking tot de variabelen in het verkeer die invloed hebben op taaklast. Deze review onderzoekt het effect van bestuurderskarakteristieken, weggeometrie, omgevingsfactoren en voertuigkarakteristieken. Onder bestuurderskarakteristieken vallen de leeftijd van de bestuurder, rijervering, geslacht, persoonlijkheid en bekendheid met de rijomgeving. Meerdere studies die leeftijd en rijervaring relateren aan taaklast zijn gevonden. Voor het geslacht van de bestuurder waren de resultaten verschillend, sommige studies vonden significante effecten maar geen consensus was gevonden. Significante effecten waren gevonden voor bekendheid met de omgeving, maar alleen wanneer de deelnemers geen navigatiesysteem gebruikt.

Drie aspecten van weggeometrie zijn behandeld: wegbreedte, kromming van de weg, en of het verkeer gescheiden is van dezelfde wegbaan gebruik maakt. Verkeer op de tegenliggende strook resulteert in een stuur demand weg van het midden van de weg, wat resulteert in een hogere taaklast. Het verminderen van de wegbreedte zorgt ook voor een hoger stuur demand, wat ook resulteert in een hogere taaklast. Grootte van de wegkromming en snelheid over de bochten resulteren ook in een verhoogde taaklast. Verder was gevonden dat het rijden over een ongecontroleerde kruising of rotonde in een hoge taaklast resulteert.

Voor omgevingsaspecten werden een aantal gebeurtenissen die zich voor kunnen vallen in de rijomgeving bestudeerd. Hierover vallen de aanwezigheid van kwetsbare verkeersdeelnemers en vrachtverkeer, het effect van veranderingen in verlichting en weer, omgevingscomplexiteit, en gedrag en aantallen van andere voortuigen op de weg. Hoewel er geen onderzoek is gevonden die het effect van de aanwezigheid van kwetsbare verkeersdeelnemers op taaklast bestudeerd, wordt vaak aangenomen dat deze een invloed heeft. Het effect van de aanwezigheid van vrachtverkeer is in een aantal studies aangetoond. Een vermindering in zichtbaarheid als een resultaat van verminderde verlichting of slecht weer resulteert in een verhoging van de taaklast. Verkeersdichtheid wordt gezien als een belangrijk aspect in een aantal studies, waarbij niet alleen de dichtheid van de gedeelde stroken maar ook die van tegenliggende stroken van belang is. Omgevingscomplexiteit wordt in dit onderzoek gezien als alle aspecten van het verkeer die een effect op taaklast kunnen hebben maar niet direct de rijtaak belemmeren. Voorbeelden hiervan zijn hoge gebouwen of advertenties langs de

wel, of ongevallen die zich op de andere rijbaan plaatsvinden. Een hogere omgevingscomplexiteit is een voorname reden waarom stedelijk verkeer correleert met een over het algemeen hogere taaklast dan verkeer buiten de bebouwde kom. Verder is de complexiteit van de voertuig interacties van belang. Hogere orde manoeuvres als ritsen, inhalen of van strook wisselen zijn extreem complex, aangezien de bestuurder beslissingen moet maken over zowel longitudinale als laterale beweging in een kort tijdsbestek, terwijl tegelijktijdig de andere voertuigen in de gaten moeten worden gehouden.

Ten slotte zijn de voertuigkarakteristieken bestudeerd, waarin vooral gekeken is naar Advanced Driver Assistance Systems (ADAS). Een aantal ADAS systemen kunnen potentieel zorgen voor een vermindering van de taaklast door bepaalde handelingen over te nemen, e.g. snelheid handhaven of navigeren. Het opereren van deze systemen tijdens het rijden kan echter wel zorgen voor een vergrote taaklast, wat in het bijzonder het geval is bij het gebruik van mobiele telefoons tijdens het rijden.

Experimentele Setup

In het voorgestelde onderzoek bekijkt de deelnemer een serie van video fragmenten, welke een combinatie van een aantal variabelen bevatten, waarvan in het verleden is gevonden dat deze een invloed hebben op mentale taaklast. Voor deze fragmenten wordt de deelnemer gevraagd om zich in te beelden in de plaats van de bestuurder. Na elk fragment geeft de deelnemer een waarde aan op een subjectieve meetschaal.

In totaal deden 60 deelnemers mee aan het onderzoek, verdeelt over 3 leeftijdscategorieën: een jonge groep (18-25), een groep van middelbare leeftijd (30-50) en een oudere groep (65+). Afgezien van de leeftijd is informatie bekend over het geslacht, rijervaring in jaren, kilometrage en de frequentie waarmee de deelnemer gebruik maakt van de snelweg. Met behulp van een software applicatie worden de video fragmenten (15 seconde lengte) een voor een aan de deelnemer weergegeven, waarna bij elke fragment een waarde op de subjectieve schaal wordt ingevuld. De subjectieve schaal die gebruikt wordt in dit experiment is de Rating Scale Mental Effort (RSME; Zijlstra, 1993), vanwege zijn uni-dimensionaliteit en gevoeligheid in verkeerssituaties waarbij prestatie optimaal is (De Waard, 1996). De deelnemer kan het onderzoek op eigen tempo uitvoeren, maar kan pas een waarde invullen nadat de volledige lengte van het fragment weergegeven is. Alle deelnemers beoordelen alle video fragmenten, welke in een willekeurige volgorde weergeven worden ten behoeve van de counterbalancing.

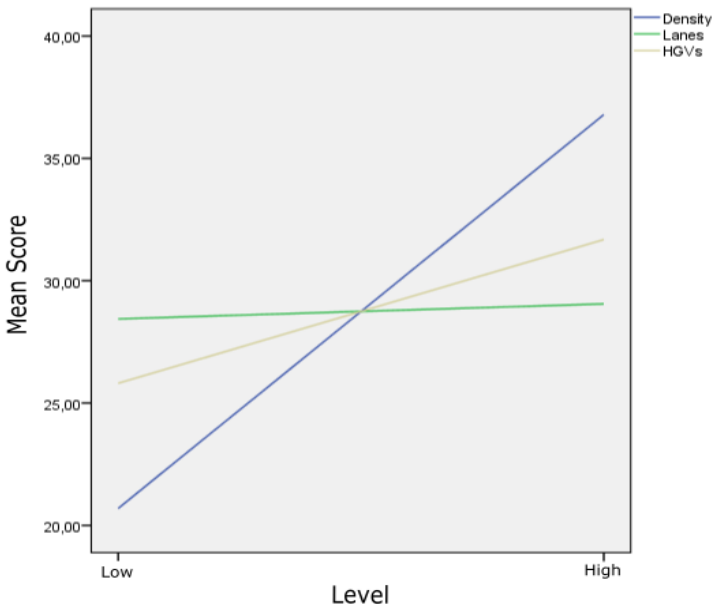
Figuur 1 - Experiment Software

Het experiment bevat twee ontwerpen. Het primaire ontwerp is een factoriaal ontwerp van de variabelen: verkeersdichtheid, aantal wegstroken en de aanwezigheid van vrachtverkeer. Hiervoor zijn alleen snelwegsituaties meegenomen, en elke combinatie van factoren is twee keer aanwezig. De reden om alleen snelwegsituaties mee te nemen is omdat deze gemakkelijk kunnen worden geuit in combinaties van een beperkt aantal variabelen en weinig beïnvloed worden door omgevingscomplexiteit, welke moeilijk te kwantificeren is door de brede scala aan bronnen van complexiteit. Het secundaire ontwerp bevat paren van vrijwel identieke situaties, met als enig verschil de aanwezigheid van een kwetsbare verkeersdeelnemer of slecht weer bij een van de fragmenten. De inclusie van dit ontwerp heeft twee doelen: het staat het onderzoeken van het effect van kwetsbare verkeersdeelnemers en slecht weer toe, en leidt de deelnemer af van het doel van het primaire ontwerp. Wanneer het experiment alleen kleine variaties van snelwegsituaties zou laten zien, is het mogelijk dat de doornemer het doel van het onderzoek doorheeft, wat invloed op zijn beslissingen kan hebben.

Voordat de deelnemer mag beginnen aan het experiment, werd een korte introductie gegeven. Het doel van deze introductie is om de procedure uit te leggen, de deelnemer bekend te maken met de schaal en om een referentieschaal te geven voor het gebruik van RSME in het experiment. Tijdens de pilot gaven deelnemers aan het moeilijk te vinden om hun eerste scores te geven, aangezien zij nog geen referentieschaal hebben opgebouwd voor zichzelf. Na een aantal video fragmenten gezien te hebben is de deelnemer beter in staat om tot een waarde te komen. Dit werd gedaan door de deelnemer drie voorbeeldsituaties te laten zien, waarbij bij elk van deze op de RSME schaal een waarde wordt aangegeven die gebruikt zou kunnen worden. De situaties die hiervoor gebruikt zijn, zijn een situatie met hoge taaklast, een met lage taaklast en een situatie die ergens tussen deze twee in zit. Na de introductie begon de deelnemer zelfstandig aan het experiment.

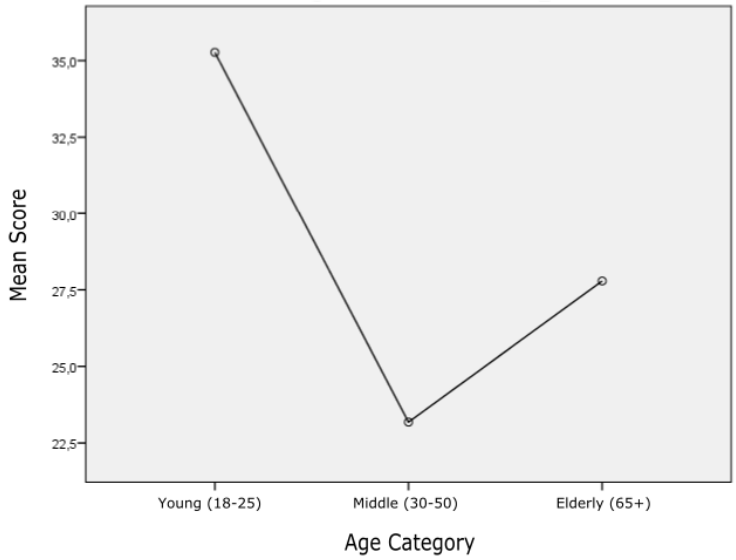
Resultaten

De resultaten van het experiment werden geanalyseerd met een Repeated Measures ANOVA (rANOVA). Significante hoofdeffecten ($p<0.05$) werden gevonden voor: verkeersdichtheid, aanwezigheid van vrachtverkeer en de leeftijd van de deelnemer. Deelnemers vonden dat de situaties met hoge verkeersdichtheid veel meer inspanning vereisten dan degenen met lage dichtheid. De aanwezigheid van vrachtverkeer vond ook een significant verschil, maar het aantal wegstroken resulteerde niet in significante resultaten. Het hoofdeffect van de within-subject variabelen is weergegeven in figuur 2.



Figuur 2 - Hoofd effect within-subject variabelen

In het experiment gaf de jongere groep aan de hoogste taaklast te ondervinden, gevolgd door de oudere groep en uiteindelijk de middelbare groep. Dit effect is uit te leggen, jongere bestuurders missen de ervaring en automatismen die aanwezig zijn in de andere groepen, terwijl oudere bestuurders een verminderde procescapaciteit hebben als gevolg van veroudering. In de meeste taaklast studies komt het echter voor dat de jongere groep zichzelf overschat, wat in dit onderzoek niet het geval is. Het hoofdeffect van de leeftijd is weergegeven in figuur 3.



Figuur 3 - Hoofd effect leeftijd

Interactie effecten zijn gevonden voor Dichtheid x Vrachtverkeer, Leeftijd x Wegstroken en Leeftijd x Dichtheid. Het effect van vrachtverkeer was het grootst in situaties met lage dichtheid. De jongere groep vond situaties met 3 stroken inspannender dan situaties met 2 stroken bij hoge dichtheid, terwijl de andere twee groepen de situaties met 2 stroken inspannender vonden. Verder gaf de jongere groep een relatief grotere toename in inspanning aan bij als een resultaat van verhoogde verkeersdichtheid. Een overzicht van de resultaten is te vinden in tabel 2.

Tabel 2 - Resultaten r-ANOVA (waarden zijn gestandaardiseerde effect groottes (η^2))

	Main Effect	Density	Lanes	HGVs	Age
Density	0.749**	x	0.001	0.162**	0.121*
Lanes	0.014	0.001	x	0.035	0.142*
HGVs	0.465**	0.162**	0.035	x	0.065
Age	0.163*	0.111*	0.142*	0.065	x

rANOVA op het secundaire design gaf een hoofdeffect van de aanwezigheid van kwetsbare verkeersdeelnemers en slecht weer aan, waarbij een interactie effect is gevonden tussen kwetsbare verkeersdeelnemers en leeftijd; hierbij gaf opnieuw de jongere groep een relatief hogere inspanning aan in de complexere situaties. Een rANOVA uitgevoerd op het primaire ontwerp waarbij het geslacht van de deelnemer was meegenomen resulteerde in een interactie effect tussen leeftijd en geslacht, maar geen hoofdeffect van geslacht. Geen significante effecten zijn gevonden als een resultaat van het toevoegen van de kilometrage.

Evaluatie van de meetmethode

Aangezien de methode die in dit onderzoek is toegepast nog ongetest is, is het belangrijk om aandacht te besteden aan de evaluatie van de methode. De methode is beoordeeld op zijn test validiteit en het experiment is beoordeel op zijn experiment validiteit. Er zijn twee oorzaken die kunnen zorgen voor een lage test validiteit. Er is maar een meetmethode gebruikt, terwijl het doorgaans normaal is om minstens twee meetmethoden te gebruiken en kunnen de resultaten van

een subjectieve meetmethode beïnvloed worden door bias, doordat deelnemer hun capaciteiten overschatten. Verder is het niet bekend wat de invloed is van het laten zien van video fragmenten, in plaats van dat de deelnemer zelf rijdt. Deze twee oorzaken kunnen elkaar versterken, aangezien wanneer van deze methode wordt gebruikt een gebrek aan capaciteit niet correleert aan een vermindering in de prestatie van de bestuurder.

Een benadering van de test validiteit wordt gemaakt door het vergelijken van de effecten die zijn gevonden met resultaten uit voorgaande onderzoeken die meetmethoden gebruiken waarvan validiteit beter bekend is. De invloed van dichtheid, vrachtverkeer en leeftijd zijn in de literatuur gevonden, welke ook zijn gevonden in dit onderzoek.

Voor de experiment validiteit zijn er een aantal mogelijke oorzaken van beperkingen in de interne en externe validiteit. Het gebrek aan beheersing over de variabelen kan zorgen voor een vermindering van de interne validiteit. Het beheersen van de variabelen was voornamelijk gestuurd op het vermijden van wegkrommingen en manoeuvres. Hierdoor kan voornamelijk de omgevingscomplexiteit variëren tussen verschillende situaties. Een serie van rANOVAs is uitgevoerd om de invloed hiervan te testen, waarbij in plaats van het gemiddelde van twee identieke situaties verschillende combinaties van enkele situaties werden gebruikt. Hieruit zijn een aantal verschillen in de significante interactie effecten gevonden, maar is het hoofdeffect van verkeersdichtheid, vrachtverkeer en leeftijd nog steeds significant in alle combinaties.

Verder zijn er ook een aantal oorzaken voor mogelijke vermindering in de externe validiteit aanwezig. Er ontbreken persoonlijke gegevens in betrekking tot enkele persoonlijke karakteristieken, zoals sociaal economische status en burgerlijke staat. Verder is er een bias richting een hoger opleidingsniveau, aangezien een groot deel van de deelnemers is gewerfd onder werknemenden en studenten van de VU universiteit in Amsterdam. Werving is gecentreerd rond de steden Den Haag en Amsterdam, wat resulteert in een hoog aandeel van deelnemers die gewend zijn om in stedelijke omgeving te rijden. De invloed hiervan is echter niet te meten zonder eenzelfde studie uit te voeren waarbij gebruik gemaakt wordt van een sample met andere karakteristieken.

Om te testen of er een bias gecreëerd is door het gebruik van een repeated measures ontwerp, zijn er een aantal tests uitgevoerd, waarbij gekeken is naar de aanwezigheid van volgorde- of leereffecten. Er zijn geen significante effecten gevonden als een resultaat van een verschil in de complexiteit van de eerste video, de volgorde en afstand tussen de gekoppelde fragmenten, of leereffecten.

Conclusie

Het doel van dit onderzoek is het ontwikkelen van een methode waarmee mentale taaklast kan worden gemeten, gebruik makend van video fragmenten, en het peilen van de bruikbaarheid van deze methode in mentale taaklast onderzoek door uitvoering van een experiment. Hiertoe zijn een aantal onderzoeksvragen ontwikkeld:

1. Wat is mentale taaklast en hoe kan het gemeten worden?
2. Welke aspecten van autorijden kunnen tot verhoogde mentale taaklast leiden?
3. Hoe kunnen video fragmenten gebruikt worden om taaklast te meten?
4. Wat is de validiteit van de meetmethode?
5. Hoe verhoudt de meetmethode zich vergeleken met andere methoden voor het meten van taaklast?

De eerste drie vragen zijn beantwoord in de eerste drie paragraven, en de vierde vraag is beantwoord in de vorige paragraaf. De laatste vraag kan beantwoord worden door te reflecteren op de meetmethode evaluatiecriteria weergegeven in tabel 1. De methode is verrassend sensitive, wat te zien is aan het grote aantal significante effecten die zijn gevonden waarbij bij een groot deel van de situaties de rijprestatie optimaal was. Aangezien van een uni-dimensionale schaal gebruik gemaakt wordt, is er geen diagnosticity aanwezig. Wanneer de primaire taak gezien wordt als de mogelijkheid van de deelnemer om zich te verplaatsen in de bestuurder van de auto, is het mogelijk dat er primary task intrusion is. Door de korte duratie en snelle opeenvolging van de videofragmenten, kan het weergeven van de schaal tussen elk fragment zorgen voor een hoge intrusie. De lage kosten en tijdsinvestering die benodigd is voor de uitvoering, zijn de voornaamste reden voor het gebruik van de methode. Door de korte tijdsbenodigdheid is de operator acceptance hoog, hoewel de deelnemer niet altijd overtuigd is van de nuttigheid van het onderzoek. Er werd vaak opgemerkt dat geen van de situaties echt inspannend was. Om de selectiviteit te bepalen moeten vergelijkingen worden gemaakt met methodes waarvan bekend is dat deze taaklast meten, en is om deze reden grotendeels onbekend. Omdat deze methode voor het eerst is toegepast voor het bepalen van taaklast, is geen informatie bekend over de bandwidth and reliability van de methode.

De methode die is ontwikkeld in dit onderzoek heeft potentie om gebruikt te kunnen worden als valide en gevoelige methode voor het meten van taaklast. De methode demonstreert zijn gevoeligheid door zijn vermogen om significante effecten te vinden bij variërende aanwezigheid van factoren. Hoewel er geen definitieve uitspraak gemaakt kan worden over de validiteit van de methode, is het toch aanbevolen om deze methode waar mogelijk te gebruiken voor het meten van taaklast. Vervolgend gebruik van de methode, vooral in combinatie met andere taaklast meetmethoden, zal de validiteit van de methode verbeteren.

Table of Contents

1 Introduction

Driving involves high fluctuations in mental workload, since a wide variety of demands is placed on the driver in a high sequence. At a high workload it is possible for the driver to become overloaded and unable to respond to new information. Low workload can result in boredom, decreased situational awareness and an overall reduction in alertness. Both high and low workload are causes of driver's inattention, which is seen as one of the primary causes of traffic accidents. In order to better prevent traffic accidents, it is important to understand the cause and mechanics behind mental workload. For this reason, workload is the subject of a large body of research. Mental workload research is generally performed using either instrumented vehicles or driving simulators. These studies are generally time-intensive and expensive to carry out. This research describes the development of a workload measure method which has the primary benefit that it is very easy and inexpensive to implement.

In this research, the influence of several traffic variables on car driver's mental workload, using a method developed in this study. With this method, short video fragments of traffic situations are shown to the participant and they are asked to rate their perceived mental effort imagining that they are driving through these situations themselves. By comparing the results obtained through this method and comparing them with results obtained in studies performed in the past, an evaluation of the measurement method can be made. The videos used for this method are derived from naturalistic driving data, data which is gathered with instrumented vehicles in naturalistic driving settings. The data is gathered in the ITS project, a project carried out at the Stichting Wetenschappelijk Onderzoek Verkeersveiligheid (SWOV), in which a natural driving setting is replicated by providing a vehicle instrumented with cameras and sensors to participants for a period of four weeks. This vehicle then returns with data from every trip that has been made during this period. The data contains speed, accelerations, location, time period and video images of the driver and front view of the car.

1.1 Relevance of the study

The concept of mental workload measurement and the effect that certain variables in traffic have on mental workload have been extensively studied in the past. In this study however a relatively untested method for obtaining mental workload is developed, letting subjects judge video images on perceived effort rather than having them drive themselves. The advantage of this method is that it is possible to examine a great number of factors, since the judgment of a single combination in the factorial design only takes a short amount of time. The disadvantage however is that because it is an untested method of obtaining workload, it is unclear how valid the results obtained in this way will be. It is however interesting to see if the effects obtained through this method match the results found in earlier studies. If it turns out the study produces a valid and reliable way to measure workload, it is a useful method to measure workload for any study using naturalistic data. The method is inexpensive and requires little time to implement while allowing the study of a lot of factors.

1.2 Research objective

The objective of this research is to develop a method through which mental workload can be measured using video fragments, and to gauge its usability in mental workload research though the use of an experimental application. In order to work towards this objective a number of research questions are created:

1. What is mental workload and how can it be measured?
2. What aspects of car driving can result in increased mental workload?
3. How can video fragments be used to measure mental workload?
4. What is the validity of the measurement method?
5. How does the measurement method compare to other workload measurement methods?

Answering all the research questions should ultimately lead to the fulfillment of the research objective. The following chapters are structured in a way to allow the answering of the individual research questions. In chapter 2, the concept of mental workload and its measurement are explained. In chapter 3 the different variables which are found in traffic are examined for the influence on mental workload that has been found in studies the past. In chapter 4 the experiment that is employed is developed and explained. In chapter 5 the results of the experiment are displayed and analyzed. In chapter 6 the research method and results are evaluated and in chapter 7 a conclusion is drawn on the results which were found in the experiment and recommendations are made for further use of the study.

2 Mental workload and measurement

The use of the term mental workload may be a difficult concept for people who are not familiar with it. This chapter gives a definition of mental workload and explains some of the terms related to mental workload.

There is no current standard definition of mental workload. Instead, different researchers use different definitions. In a review of mental workload literature, Cain (2007) mentioned that the formal definitions researchers use are all slightly different, but however share some common traits. In this report the definition of mental workload used by De Waard (1996) is used. Here, mental workload is defined as the specification of the amount of information processing capacity that is used for task performance. A person's information processing capacity is not only determined by their skill and experience in performance of the task, but also by their capability in performing the task. This capacity can change with an increasing task demand when a mental effort is made by the operator. Task demand is the processing requirement to perform a task with desired performance, independent of the capability of the individual performing the task. Task demand is largely determined by the complexity of the task, which is reflected in the number of processing stages that are required to perform the task. Task difficulty is the processing effort, specific to an individual, which is required for the task and is determined by factors such as task demand, processing capacity, experience and state of mind. It is important to note that task capabilities of an individual are not constant in every situation. Increasing demand does not necessarily result in an increase in workload, since capabilities may be low due to the effect of, for instance, boredom. Going from a state of boredom to a state of concentration by investing mental effort will often decrease mental workload.

Mental workload is considered one of the primary indicators of driving safety. Drivers are more likely to enter risky situations when demand placed upon them exceeds their capacity (Wong, 2009). When a driver is unable to process the information inflow, they may miss something important or lose control over the vehicle, which often results in traffic accidents. Inattentional blindness is defined as the inability of a person to see an unexpected stimuli appearing in their field of vision, and is the cause of many traffic accidents. This is often caused by a cognitive overload resulting from the driver being too busy with other tasks, such as talking on the phone or watching road-side advertisements, which can be a result of either high or low demands being placed upon the driver. An increase in mental workload shows high similarity with- and is often accompanied with an increase in driver distraction (Schaap et al., 2009). By increasing knowledge on the source of a high workload, it is possible to create preventive measures and ultimately reduce the occurrence of traffic accidents. For this reason, many researchers focus on the measurement of mental workload in different situations. As a result a large number of measurement methods has been developed over the years, ranging from measuring their heart rate, to simply asking the participant for their perceived workload. In the following paragraphs these measures are explained, as well as the quality criteria which are used to evaluate the strength and weaknesses of the measurement methods.

2.1 Workload and task performance

In order to be able to evaluate different mental workload measurement methods it is important to understand the relationship between mental workload and task performance. Although not originally

described as such, Yerkes and Dodson (1908) found the basis for a relationship between arousal and performance, which was later developed to what is known as the Yerkes-Dodson law. The Yerkes-Dodson is illustrated graphically as an inverted-U curve. The concept of the inverted-U function is that the performance of a task increases when arousal increases, up to a certain point after which the performance diminishes. Figure 2.1 shows an illustration of the inverted-U curve for driver performance (De Waard, 1996) in which arousal is substituted for task demand.

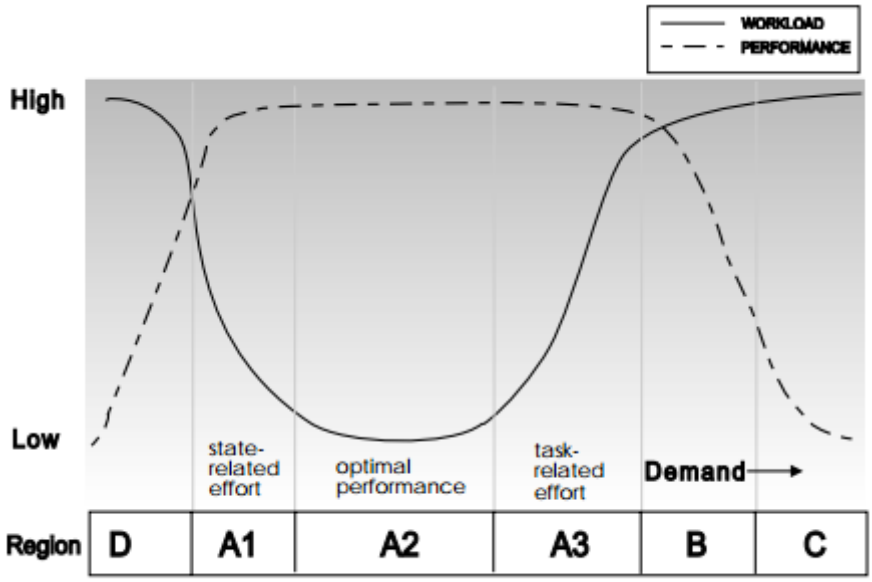


Figure 2.1 - Inverted U-curve of workload and performance (De Waard, 1996)

In Figure 2.1 the performance and workload are shown over an increase in task demand according to the Yerkes-Dodson law. In region A2 the operator can easily deal with the task demands and reach their preferred level of performance. As a result, performance is optimal and workload is low. When demand decreases, a state related effort needs to be made in order to maintain performance by counteracting the capacity loss due to inattention. When the demand is too low for effort to have an effect, region D is entered and performance decreases. When demand increases from region A2 a task related effort needs to be made in order to maintain optimal performance. When the demand is too high to maintain performance region B is entered. When performance has lowered to a minimum level, region C is reached. It is possible that the cognitive demand is so high, the operator 'gives up' on the task. When this is the case a reduction in the workload can be measured, however performance remains minimal.

2.2 Measurement Quality Criteria

A number of criteria which are used for the evaluation of mental workload measurement methods is provided in O'Donnell and Eggemeier (1986), which was subsequently expanded with a number of criteria in De Waard (1996). In the development of the measurement method it is important to understand the criteria which determine the quality of the method and which will eventually be used to evaluate on the method.

2.2.1 Sensitivity

Sensitivity is determined by the extent to which the measurement method can detect changes in the level of mental workload related to the task performance. In paragraph 2.1 the relation between mental workload and performance was explained. In different parts of the inverted-U model a measurement method may have a different sensitivity to variations in the level of mental workload. For this reason it is important to employ multiple methods of workload measurement methods, since different methods may be sensitive to different areas in the inverted-U model.

2.2.2 Diagnosticity

Diagnosticity relates to the multiple resource theory proposed by Wickens (1991). According to the theory, a human operator has multiple processing sources from which can be drawn simultaneously. Depending on the nature of the task, different pools of resources can be utilized. As suggested by Wickens (1991), the perceptual, central processing and motor input stages draw from different pools and could therefore be performed at the same time without interference. Diagnosticity is the capability of the assessment method to discriminate the workload on the different available resource pools.

2.2.3 Primary task intrusion

This criterion refers to the intrusion that the measurement method has on the performance of the primary task. In measurements in which a secondary task is added to the primary driving tasks, it is important that the primary tasks performance does not degrade as a result of the addition of the secondary task. Since the effort is measured by judging the secondary task performance, when the secondary tasks uses resources that would normally be attributed to the primary task the results are not accurate.

2.2.4 Implementation requirements

Implementation requirements concern the constraints in the execution of the measurement method. These include the availability of equipment and instruments, time investment and cost. While this criterion is can be viewed as less important than the aforementioned criteria, the method developed in this study is focused mainly on having low implementation requirements.

2.2.5 Operator acceptance

The operator acceptance refers to the willingness of the subject to participate in the experiment and their perception of the validity and utility of the procedures. This criteria is important since this can determine the effort the participant is willing to put into the experiment. When the participant does not believe the experiment to be valid they may put in a less than optimal performance.

2.2.6 Selectivity

Selectivity is the degree to which the measurement method is sensitive to specifically the trait that is being researched. For mental workload this means that measurement methods that have good selectivity are mostly sensitive to the mental workload of the subject, and not to for instance physical load. A measurement method with poor selectivity can still be used, however the circumstances will have to be controlled to such extend that the other traits to which the method is sensitive are possible to filter out either during the experiment or the processing of the results.

2.2.7 Bandwidth and reliability

Bandwidth and reliability refer to the reliability of the measurement method both within and across tests. A method with good bandwidth is applicable in different situations and different performance regions in the inverted U-model, while a method with good reliability will produce results that are consistent in repeated measurements.

2.3 Measuring mental workload

In this paragraph the different methods which are used to measure mental workload are discussed. O'Donnell and Eggemeier (1986) identified four different groups of workload measurement methods. These are subjective measures, primary and secondary task performance measures, and physiological measures. For each of these groups, some examples of tests which resort under the respective category are provided.

2.3.1 Subjective measures

Subjective measures are performed by having the participant rate their perceived effort or complexity of the task, after performing the experimental task. There are a number of rating scales, which are widely used in subjective workload research. These scales can be divided into uni-dimension and multi-dimensional scales (De Waard, 1996). For uni-dimensional scales the subjects indicate their perceived effort or activation on a single scale, while for multidimensional scales the subjects need to complete multiple scales, such as mental demand, frustration level and effort.

Three uni-dimensional scales are mentioned in De Waard (1996): the Rating Scale Mental Effort (RSME; Zijlstra, 1993), the activation scale and the Modified Cooper-Harper scale (MCH; Wierwille and Casali, 1983). Conceptually, the RSME and activation scales are similar; they both consist of a single line along which the subjects indicate their perceived mental effort, with anchors along the line that serve as a reference. The main difference in design is the nature of the descriptions of the anchors; the RSME gives a direct indication of subjective effort (low effort, moderate effort, very high effort etc.), while the activation scale shows examples of tasks of a comparative effort level (solving a crossword puzzle, trying to cross a busy street etc.). Furthermore, on the activation scale, participants are asked to indicate their mental activation, rather than their mental effort. The MCH follows a flow diagram, in which a series of questions leads to a rating for mental workload from 1-10. Since uni-dimensional scales do not account for the different aspects of workload, an uni-dimensional scale has by definition no diagnosticity.

Two multi-dimensional measurement methods were mentioned in De Waard (1996): the NASA Task Load Index (TLX; Hart and Staveland 1988) and the Subjective Workload Assessment Technique (SWAT; Reid and Nygren, 1981). In these measures, ratings for different subscales have to be provided by the subjects, such as ratings for physical demand, frustration level and performance. This results in a higher level of diagnosticity. However, it can be assumed that these multi-dimensional scales may make it complicated for the subjects to identify the different aspects of workload. Both scales have a similar sensitivity, however the TLX seems to have a higher operator acceptance (Cain, 2007).

There are several advantages to using subjective measures: they are relative easy to implement and read out and have low to no primary task intrusion. While the subjective nature may make these measures seem susceptible to bias, multiple researches suggest that self-report measures are both

valid and sensitive indicators of workload (Gartner and Murphy, 1976; Johannsen et al, 1979; Sheridan, 1980, as seen in De Waard, 1996). Other disadvantages of self-report measures are differences in participant's preference in placing their ratings on the scale, as well as biases caused by overestimation of the subject's own capabilities.

A person's personality plays a role in their mental workload. The personality of a driver can be broken down into multiple aspects. Examples that are relevant in traffic are: sensation seeking, altruism, normlessness, anger and impatience (Wong, 2009). Wong et al. (2010) suggested that personality traits can influence the occurrence of risky driving behavior, which results in higher task demands. A sensation seeking driver will often drive faster than a normal driver, which increases the inflow of information and thus requires greater information processing capacity.

2.3.2 Primary task performance measures

Primary task performance measures use several aspects of the primary task performance as an indicator of mental workload. Common examples are lateral movement and speed maintenance. Primary task performance measures are most sensitive when task demand exceeds capabilities, resulting in a reduction of performance, and is very insensitive to changes in workload at an optimum performance level. During performance of the driving task, performance will most of the time be in the optimal performance region of the inverted U-model. This means that even though there may be a difference in workload, performance is generally stable, making this measurement method insensitive in this region (De Waard, 1996). For this reason primary task performance measures are generally coupled with other measurement methods when determining mental workload.

2.3.3 Secondary task performance measures

In secondary task performance measures an additional task is added to the primary task. Performance aspects of the secondary task, such as reaction time or number of errors, are used as a measure. Examples of secondary tasks that are often applied are the Peripheral Detection Tasks (PDT) mental calculation tasks or digit recall tasks. The advantage of using secondary task measures is that they can be used continuously during the tasks and therefore provide insight into certain scenarios during the experiment. A nonintrusive self-report measure will only provide insight into the entire task and it is unclear whether the overall task load or peaks are indicated (Martens & Van Winsum, 2000). The usefulness of the secondary task performance measure depends on a two factors: the addition of the secondary task should draw from the resource pool that is concurrently used by the primary task in order to obtain a reduction in secondary task performance and the secondary task should not intrude on the performance of the primary task (De Waard, 1996). The first conditions makes secondary task performance measures especially sensitive when a performance related effort needs to be made by the driver. De Waard argued that the primary task is less intrusive when the secondary task is not artificial but integrated in the driving tasks, examples of this are rear-view mirror checking and radio tuning.

2.3.4 Physiological measures

Physiological measures are observations of the operator's state, through measures of physiological processes. Commonly used methods are Electroencephalography (EEG), Electrocardiography (ECG), Electromyography (EMG) and eye activity. While these methods suffer from obtrusiveness, with current technological advances the equipment is steadily becoming more practical in use and has the potential to be an unobtrusive, objective measurement method (Cain, 2007).

A disadvantage of a lot of physiological measures is that they have a low selectivity. The result of methods measuring heart rate, heart rate variability or blood pressure do not distinguish between physical- and mental load. Furthermore eye activity measures are sensitive to fatigue and time on task (Stern et al., 1994; Fukuda et al., 2005). This results in the fact that the validity of the measures is determined by the skill of the user in monitoring and control over the outcome. This means that most physiological measures require expertise in the reading out of the results. When properly controlled however, physiological measures tend to be a very sensitive and a valid method of determining workload.

2.4 Chapter Summary

This chapter discussed the concept of mental workload and its related terms and as such is an attempt to answer the first research question: What is mental workload and how can it be measured? While there is no agreed upon definition of workload which is used by all researchers, there are some common traits. Overall, mental workload is described as the relative information processing capacity that is employed to achieve task performance. Two factors are important in the determination of workload: the demand of the task and the skill and experience of the person performing the task. Task demand is a processing requirement to perform a task with desired performance, independent of the capability of the individual performing the task. Task demand is largely determined by the complexity of the task, which is reflected in the number of processing stages that are required to perform the task. Task difficulty is the processing effort, specific for an individual, that is required for the task and is determined by factors such as task demand, processing capacity, experience and state of mind.

The relationship between workload and performance can be pictured with an inverted U-curve over a u-curve (see Figure 2.1). At low task demand, performance is low and workload is high as a result of boredom. When demand increases a state-related effort is made and performance increases while workload decreases. When task demands increase further, a performance related effort needs to be invested to maintain performance. This cause workload to increase while performance remains stable, up to a point where the demand is too great which causes a reduction in performance.

A number of measurement evaluation criteria were mentioned in O'Donnel and Eggemeier (1986) and De Waard (1996). Sensitivity is determined by the extent to which the measurement method can detect changes in the level of mental workload related to the level of task performance. Diagnosticity relates to the ability of the measurement method to discriminate the workload from different resource pools, which are the perceptual, central processing and motor input (Wickens, 1991). Primary task intrusion refers to the degree to which the measurement method intrudes upon the primary task, which is safety operating the vehicle in most driving workload research. Implementation requirements are the expertise, time and financial requirements that are necessary to perform the measurement method. Operator acceptance refers to the willingness of the subject to participate in the experiment and to what degree they think the research is valid. Selectivity is the degree to which the measurement method is sensitive to the specific trait that is being researched. Finally, bandwidth and reliability refer to the consistency and applicability of the method over different repeated measures and applications on different areas of the U-model.

Four different categories of measurement methods were discussed. These are subjective measures, primary- and secondary task performance measures, and physiological measures. When using

subjective measures, the participant is asked to rate their perceived effort after completing the task. These measures generally have a high sensitivity and low primary task intrusion, but can be subject to bias. Primary task performance measures use measures in task performance such as lane keeping or speed maintenance to determine workload. These measures are generally insensitive to changes in workload during optimal performance. In secondary task performance an additional task is included, of which common examples are calculation or reaction tasks. A reduction in secondary task performance could be measured during high workload situations, where more attention needs to be directed towards the primary task. For this it is important that the secondary task does not intrude upon the primary task, so only the secondary and not the primary task performance degrades during these situations. Physiological measures are observations of the operator's state, through measures of physiological processes. Commonly occurring examples are the use of skeletal muscle or heart monitoring devices. It is important that a good control over the physical activity of the subject is maintained however, since these measures are often sensitive to this.

Table 2.1 displays the different measurement method categories and their strengths and weaknesses. Quality of the different methods in the table are indicated by a +, – or 0. A + indicates a positive index, 0 an average index and – a negative index. Indexation is done on a relative level, e.g. subjective and secondary task performance score 0 at sensitivity although they are considered generally sensitive, since physiological measures have such high sensitivity.

Table 2.1 - Measurement Method Evaluations

	Sensitivity	Diagnosticity	Prim. task intrusion	Implement. req.	Operator acceptance	Selectivity	Bandw. & reliability
Subjective	0	- / 0*	+	+	+	+	0
Prim. task	-	-	+	+	+	+	0
Sec. Task	0	+	-	0	0	0	+
Physiological	+	+	0	-	0	-	+

**- for uni-dimensional scales and 0 for multi-dimensional scales*

3 Variables in traffic affecting mental workload

In an attempt to gain a better understanding of mental workload, aspects of traffic that may have an impact on workload have often been studied. This chapter provides a look at the relevant literature. The variables that are considered in this chapter are the prominent variables that may vary in different traffic situations and for which a possible effect on mental workload can be found. The different variables are classified into four categories: driver characteristics, road geometry, environmental factors and vehicle characteristics. For each of these categories, the individual variables belonging to this category and their impact on mental workload is explored.

The purpose of this chapter is to gain insight into the factors that determine workload, so that a decision can be made on the variables that are studied using the proposed measurement method. To this end. a great number of variables is studied. Based on the findings in this chapter, a selection is made for the variables to be studied in this research. In the end the results found from past studies can then be compared to the result from this study, as a means of evaluating the method.

3.1 Driver characteristics

The driver characteristics are the attributes of the driver which may have an effect on mental workload. The characteristics that are examined in this analysis are driver age, driving experience, gender and familiarity with the surrounding.

Age is by many researchers considered to be one of, if not, the most important aspect in mental workload for automobile drivers. Aging causes a reduction in older drivers' information processing in perceptual, cognitive and psychomotor aspects; resulting in them needing a longer time to process the sensatory information that is propelled at them during driving (Wu & Liu, 2007).

In research on the effect of age, generally 3 groups are distinguished: younger drivers (20-30), middle-aged drivers (40-50) and older drivers (65+). The largest differences can be found when comparing the younger or middle aged drivers to the older driver, where younger and middle aged drivers tend to show similar results (Makishita and Matsunaga, 2006). Older drivers have been found to show higher response times in secondary performance tasks when compared to younger drivers. However, this effect is mostly apparent in high complexity situations and tends to scale disproportionately with the increase in complexity resulting from more complex driving situations (Verwey 2000; Cantin et al., 2009). Most research performed on the effect of age on mental workload uses a form of secondary task performance measures. Using subjective measures to determine the workload of older drivers may not necessarily result in a higher workload, since older drivers have a tendency to overestimate their task capabilities, even when their performance was noticeably worse (De Waard et al., 2009; Freund et al., 2005).

Driving experience can be expressed in two quantities: the amount of years the person is in possession of a driver's license or the number of kilometers a person drives over a period of time, with experimental setups using either one, or a combination of both. Much study on driving experience has been performed on visual search patterns during driving. Novice drivers tend to have narrower visual search pattern, and make less glances on the rearview mirrors (Crundall and Underwood., 1998; Mourant & Rockwell., 1972). An explanation that could be provided for this is that the processing effort required for processing all sensatory data is higher for novice drivers. In an

experiment, in which 40 professional drivers and 39 novice drivers were compared using a Peripheral Detection Task (PDT), Patten et al. (2006) found a significant difference in reaction time between novice and professional drivers. The results show a significant difference between medium and high complexity situations for the novice drivers, but an insignificant difference between low and medium complexity situations for the professional drivers, which suggests that the more experienced drivers could remain in automated driving behaviour at a relatively higher complexity.

Gender of the participants is sometimes included as a between subjects variable. Generally no significant main effect is however found (Green et al., 1994; Teh et al., 2014) however an interaction effect can often be found with age when an elderly age category is used (e.g. Green et al., 1994) since older males' health tends to deteriorate at an earlier age than older females'.

Navigating through an unfamiliar area increases workload considerably (Verwey and Janssen, 1989; Parkes et al., 1991), which is likely caused by the need to handle maps while driving, retaining route information and finding their own location. Verwey (2000) examined the influence of familiarity with the surrounding area in drivers' mental workload. In an experiment two groups of participants drove the same route, of which one group indicated familiarity with the surroundings. The results showed no significant effect on secondary task performance however, but this was attributed to the use of a support system that provided guidance instructions. In the context of this study, since the fragments require no navigational tasks from the subject this means that familiarity with the surrounding is likely not an issue.

3.2 Road Geometry

In road geometry two types of roads can be distinguished, single- and dual-carriageway roads. The difference between the two is that in single carriageway roads opposing traffic directions drive on the same road, resulting in great speed differences between vehicles driving relative close to each other. Simulator studies often include traffic on opposing lanes as a means to increase workload. Drivers have a tendency to steer towards the edge of the road when meeting a vehicle on the opposing lane (Rosey et al., 2009) which introduces an increased steering demand.

While it can be assumed to be an important aspect of road geometry, no research could be found on the influence of the number of lanes on mental workload. However, some studies were found on the impact of lane width on mental workload. Green et al. (1994) found a main effect of the influence of road width on mental workload, with decreasing workloads as the road became wider. However the placement and number of curvatures of the roads with different widths were not the same, which in combination with the lack of counterbalancing resulted in a higher average rating for the 24 feet road compared to the 22 feet road. In Dijksterhuis et al. (2011) a main effect was found for the road width on single carriageway roads, with also an interaction effect with the density of oncoming traffic. An increase road width resulted in a decreased self-reported workload, i.e., a small decrease at low density but a larger decrease at a higher densities. The cause for the increased workload in a decreased lane width is suggested to be the result of an increased steering demand (Godley et al., 2004; Dijksterhuis et al., 2011).

In Verwey (2000) participants navigated a route in which different traffic situations were identified and compared. In this research, standing still at traffic light, straight ahead at outer- and inner-city roads, curve driving, roundabout driving, motorway driving, and straight ahead and turning at

controlled and uncontrolled intersections was investigated. Based on the results of the experiment, an index was created grouping the different maneuvers in relation to the performance on two secondary tasks: visual detection and auditory calculation task. Driving straight ahead was determined to be of low effort; curve driving and turning at controlled intersections of medium effort and turning at uncontrolled intersection and roundabouts was considered to be of high effort. In Hancock (1990), mental workload for left-turning, right-turning and straight driving on intersections was investigated. In the research a secondary response task and two subjective scales (TLX and SWAT) were used. Furthermore s the head-reversal frequency and eye-blink frequency were examined. The results showed a significant difference between the two turning maneuvers on the one hand and driving straight on the other hand for all measurement methods, but no significant difference between turning left and turning right on the intersection. Heger (1998) studied the influence of curvature on mental workload. EMG and subjective measures showed an increase in workload as a result of both higher curvature change rate as well as high speed during curves.

3.3 Environmental factors

Under environmental factors, anything that has a chance of occurring in the traffic environment which is not an aspect of road geometry is included. Among this are the presence of Vulnerable Road Users (VRUs) and Heavy Goods Vehicles (HGVs), effects of changes in lighting and weather, environmental complexity, and behaviour and number of other vehicles on the road.

Little quantitative study has been performed on the effect of the presence VRUs on driver's mental workload. However, placing vulnerable road users, most often pedestrians on the sidewalk, in simulated surroundings of traffic situations are often used to increase the complexity of a situation (Cantin et al., 2009; De Craen et al., 2007; Edquist et al., 2012; Paxion et al., 2013). The effect of the presence of Heavy Goods Vehicles (HGVs) was studied in De Waard et al. (2008, 2009). The results showed an increase in subjective workload and a change in primary performance indicators speed, speed SD, lateral position and lateral SD when HGVs were present during merging. While the presence of HGVs during the more complex maneuver of highway merging was studied, it is not clear if their presence could have an effect on mental workload in less complex situations. Differences found between self-reported effort for joining and exiting the lane indicate that the influence of HGV is likely to be larger with an increased complexity.

Hogema and Veltman (2002) found, through performance, physiological and subjective measurement, an increase in workload in situations without lighting compared to situations with lighting. The situations that were compared were driving in darkness with or without street lighting. Konstantopoulos et al (2010) found a reduction in eye movements during nighttime driving, as well as during rain. The effect of reduced visibility due to fog is sometimes added as an additional factor to workload studies; in simulator studies this can be easily administered by reducing the visibility range on the monitors. Based on the results of these experiments, a main effect of visibility is found in some studies (e.g. Brookhuis et al., 2009) but not in all of them (e.g. De Waard et al., 2008; Hoogendoorn, 2012)) based on self-reported workload.

A number of studies include traffic density as an independent variable in mental workload research, and found that increased density leads to a higher workload (e.g. Dingus et al., 1989; Brookhuis et al., 1991; Zeitlin, 2005; De Waard et al., 2008). In Teh et al. (2014) the influence of traffic flow and presence of lane changes of other vehicles on the road is examined. The results showed a main effect

of both increasing traffic flow as well as the presence of lane changes. Besides density of the lanes shared with traffic in the same direction, an effect on subjective workload has also been found of the density of oncoming traffic (Dijksterhuis et al., 2011). This is explained by indicating that the steering demand increases as a result of a participant's tendency to move towards the edge of the road, which results in higher workload.

In this report, environment complexity is considered to be a result of distractions placed in the drivers' surroundings, which do not directly interfere with the driving task. An obvious example is advertisement billboards along the road, but examples such as the presence of traffic signs or nearby buildings are also included. In a simulator study, Horberry et al. (2006) examined the effect of environment complexity on mental workload; using primary performance measures mean speed and deviation from speed limit, a secondary task in the form of a hazard detection task and subjective measures (NASA TLX). Increased environment complexity was simulated by adding billboards and advertisements and buildings, oncoming traffic and other highway furniture. Hoogendoorn et al. (2010) studied the effect on mental workload of incidents occurring on the other driving lane. Physiological and primary task performance measures showed an increase in workload, however subjective measures showed no significant difference.

Harms (1991) performed an experiment in which participants drove through a rural and an urban area, during which they performed calculations tasks. Mental workload was measured through reaction times and also through variations in speed with a 100 meter interval. The results showed an increased in the overall calculation time during urban driving. Furthermore the driving speed was lower at the intervals where the calculation time was high in urban driving. Zeitlin (1995) performed an experiment using truck drivers driving between downtown Manhattan and upstate New York over a four year period. Results from secondary performance tasks and subjective workload showed a significantly higher workload in downtown traffic.

An important property of urban traffic is the frequent opportunity of on-street parking. On-street parking results in a lower road width and causes a tendency for drivers to driver closer to the center of the road. Street-side parked cars have the disadvantage that they obstruct the view of the driver, may conceal crossing pedestrians and can suddenly become moving cars. Edquist et al. (2012) examined the effect of on-street parking cars on the driver's mental workload. Situations with no roadside parking spaces, empty road-side parking spaces and filled spaces are compared using self-report (NASA-TLX), primary task performance (lateral position and its standard deviations) and secondary task performance (PDT) measures. Furthermore critical situations were added in which a crash with a crossing pedestrian needed to be avoided. The results of the experiment showed significant differences in the lateral movement, PDT response times, TLX ratings and reaction times to pedestrians between the situations in which there were no parking spaces and filled road-side parking spaces.

Driving maneuvers of a higher order than speed maintenance and lane keeping tend to have a much higher mental demand. Higher order tasks are tasks performed on a tactical level, which require the performance of multiple tasks on an operational level (Michon, 1985). Examples of these are overtaking, merging and lane changing. Merging into traffic is a very complex task, since in a short period of time decisions need to be made on lane changing, acceleration and deceleration (De Waard et al., 2008). Through a survey, Hills and Boyle (2007) found participants find merging into heavy

traffic to be one of the most stressful aspects of the driving task. Merging, overtaking and lane changing are maneuvers similar in concept and are associated with high complexity, as they involve both the control over longitudinal as well as lateral dynamics while having to be aware of other vehicles on several other lanes (Cantin et al., 2009; Habenicht et al., 2011).

3.4 Vehicle characteristics

In this paragraph the characteristics of the vehicle are discussed. Most important are the various advanced driver assistance systems (ADAS), such as automated cruise control, lane keeping assistance and navigation devices. These systems can potentially reduce the difficulty of the driving task by taking over a number of tasks the driver would normally have to perform. However, they may also make the driving task more complex as a result of an increasing amount of input requirements and warnings provided by the systems. Due to the ever increasing amount of innovation of in-vehicle technologies, research on the behavioural and workload effects of these technologies is also increasing substantially. This paragraph discussed only a few examples of in-vehicle technologies, since a full overview of all innovations in this field is beyond the scope of this research.

By taking manual speed maintenance out of the equation, ACC has the potential to reduce the driver's mental workload. Some studies showed no significant effect (Nilsson, 1995; Ward et al., 1995; as seen in Young et al., 2004) while others found a significantly lower effort when driving with ACC (Hoedemaeker and Brookhuis, 1998; Ma and Kaber, 2005; Vollrath et al., 2011). Young et al. (2004) suggested that the implication on workload depends on the nature of the task. In a straight road driving environment with only fluctuations in traffic flow a significant effect can be found however when more demanding tasks, such as steering, are included the inclusion of ACC does not have a significant effect.

The use of a GPS-based navigation devices potentially has a positive effect on reducing the driver's workload. Through route guidance by the device, the driver has to worry less about the effects caused by unfamiliarity of the surroundings. Verwey (2000) attributed the difference in findings for familiarity with the surroundings between the 1993 and 2000 research to the use of a support system providing guidance instructions. It is possible however that through the introduction of an additional visual distraction the workload increases. Especially when the device is operated during driving a high increase in workload can be found (Green, 2004).

While not technically an in-vehicle technology, use of the mobile phone is also included in this paragraph, as it shows similarities with operating other ADAS and is very often used during driving. Talking on the phone increases workload of the driver for both handheld as well as hands-free conversations, and as a compensatory effort, drivers tend to reduce their speed and increase following distance (Brookhuis et al., 1991; Alm and Nilsson, 1994; Strayer, Drews and Johnston, 2003; Törnros and Bolling, 2006). Between handheld and hands-free conversations, no difference was found for subjective or secondary task performance measures. However, a difference in driver behaviour can be found. Drivers show a significant increase in lateral movement as well as a stronger reduction in speed when a handheld conversation is held compared to a hands-free conversation (Brookhuis et al., 1991; Patten et al., 2004).

3.5 Chapter summary

In this chapter different aspects of traffic were examined with regard to their influence on mental workload in an attempt to answer the second research question: What aspects of car driving can result in increased mental workload? This was achieved by performing a review of the relevant literature. The results showed a large number of variables which may have an impact on workload, and therefore may be used in this experiment. Different variables were categorized into four categories: driver characteristics, road geometry, environmental factors and vehicle characteristics. Among driver characteristics the driver's age, driving experience, gender and familiarity with the surroundings were studied. In the discussed studies, a relation between the driver's age and driving experience with mental workload was found. For the driver's gender the results were unclear, some studies reported on significant effects, however no consensus was found. Furthermore, in a study, significant effects were found for unfamiliarity with the surroundings, but only when the subject was not using a route-guidance system.

Three aspects of road geometry were studied: lane width, road curvature and whether the road is a single- or dual carriageway road. Previous studies found that traffic on the opposing lane results in an increased steering demand away from the center of the road, resulting in higher workload. Having a decreased road width also results in a higher steering demand, also resulting in higher workload. Size of the curvature and speed over the curve is also resulting in a high workload. Furthermore it was found that driving over an uncontrolled intersection and roundabout resulted in a high workload.

For environmental aspects, a number of aspects which may occur in the driving environment were studied. Among this are the presence of Vulnerable Road Users (VRUs) and Heavy Goods Vehicles (HGVs), effects of changes in lighting and weather, environment complexity, and behaviour and number of other vehicles on the road. While no previous research studying the effect of the presence of VRUs was found, a number of studies assumed their presence results in an increase in workload. For the presence of HGVs a couple of studies showed their effect on workload. A reduction in lighting and poor weather results in reduced vision from the driver, which results in higher workload. Traffic density has been found to be an important aspect of traffic in a number of studies; not only the density of the shared lanes, also the density of the opposing lanes was determined to have an effect of workload. In this report, environmental complexity is determined as all the aspect of the traffic surrounding that could have an effect on workload but do not directly interfere with the driving task. Examples are tall buildings, presence of roadside advertisement billboards and accidents happening on the other driving lane. Increased environmental complexity is a reason why urban traffic correlates to an overall higher workload requirement, along with the increase complexity of the vehicle interactions. Higher order driving maneuvers are considered one of the most demanding aspects of driving. Examples such as merging, overtaking and lane changing are very complex, since the driver needs to make decisions on both longitudinal as well as lateral control in in a short period of time, while maintaining awareness of surrounding vehicles.

The last category which was studied is the vehicle characteristics, in which the most attention was paid to advanced driver assistance systems (ADAS). A number of ADAS systems have the potential to reduce driver workload by taking over certain aspects of the driving task, e.g. speed maintenance or navigating. Operating the device while driving can substantially increase workload however. This is especially the case with the use of mobile phones during driving.

The literature review performed in this chapter resulted in a large amount of variables which can potentially be used in the experiment. Because of the high amount of variables found to have an effect on workload, not every variable can be used in this research. In the next chapter a selection is made of the variables which will be used in the experiment, based on the descriptions made in this chapter.

4 Experimental Setup

In the proposed research, the participant watches a series of video fragments, each containing a combination of a number of variables of which can be assumed, based on prior research, to have an influence on mental workload. For these fragments the participant is asked to place themselves into the car driver's position. After each video fragment the participant indicates a rating on a subjective rating scale. In this chapter this concept is explained in greater detail. A measurement scale is selected from the subjective scales explained in paragraph 2.3.1, a description of the participants is provided and the experimental procedure is explained.

4.1 Participants

In the recruitment of the participants, the only between subject factor used as a selection criteria is the age of the participant. Three age groups were identified: a young group (18 to 25 years old), a middle aged group (30 to 50) and an elderly group (above 65). Other attributes that were gathered are the participants' gender, driving experience in years, the number of times the person makes use of the motor way each month and, if available, their yearly kilometrage. The number of times the participant uses the motor way was also included, and allowed only participants for which this value was at least 2. This ensures that people were still actively driving in the situations which are primarily studied. For each of the age groups, 20 participants were recruited, summing up to a total of 60 participants. The number of participants per group was based on the fact that similar group sizes are often used in methods which include self-report measurements (e.g. De Waard et al., 2008,2009; Brookhuis et al., 2009).

4.2 Materials

In paragraph 2.4.1 the different self-report scales that are mentioned in O'Donnell and Eggemeier (1986) and De Waard (1996) were discussed. The first choice to make is between an uni-dimensional and a multi-dimensional scales. For this experiment this choice is not difficult; multi-dimensional scales such as the TLX and SWAT require ratings for physical load and frustration levels, attributes that are unlikely to occur in the experiment. Therefore the choice is limited to an uni-dimensional scale.

A Cooper-Harper scale modified for driving is not suitable for this research, since for this method it is important for the participant to be the person performing the task. This leaves the choice between the activation scale and RSME scale. The RSME has proven to be more sensitive in detecting state-related and task-related effort, while the activation scale is more sensitive to the deactivation region and the regions where performance is affected. Since during the driving task the performance is expected to be (near) optimal, the RSME is more sensitive and therefore chosen to be used in the experiment.

Using an RSME for this experiment has a number of advantages: they are easy to administer to large groups, they require no expertise to use and the RSME has in the past proven to be a valid and sensitive measurement method for mental workload (Gartner and Murphy, 1976; Johannsen et al, 1979; Sheridan, 1980, as seen in De Waard, 1996). A disadvantage of using an RSME is that for non-extreme situations, people seem to prefer to choose a rating from which they hardly deviate in

subsequent evaluations. This value is often different for different people, resulting in a wide range of ratings for a single traffic situation from different participants and a small range of ratings for different traffic situations from individual participants; while the opposite effect is desirable. This effect may be even more apparent in the current experiment form, considering the participants are not the people performing the driving task.

One way to possibly diversify the ratings is to cut off the top part of the RSME. In all of the video fragments the performance of the driving task is optimal, so in the inverted-u model it will remain in the optimal performance region. Therefore it is not expected that any of the participants will find the effort to be extreme or higher. The decision is made to cut off the upper 30 mm of the RSME and scale up the remaining part, giving the utilized range of the RSME image a larger display.

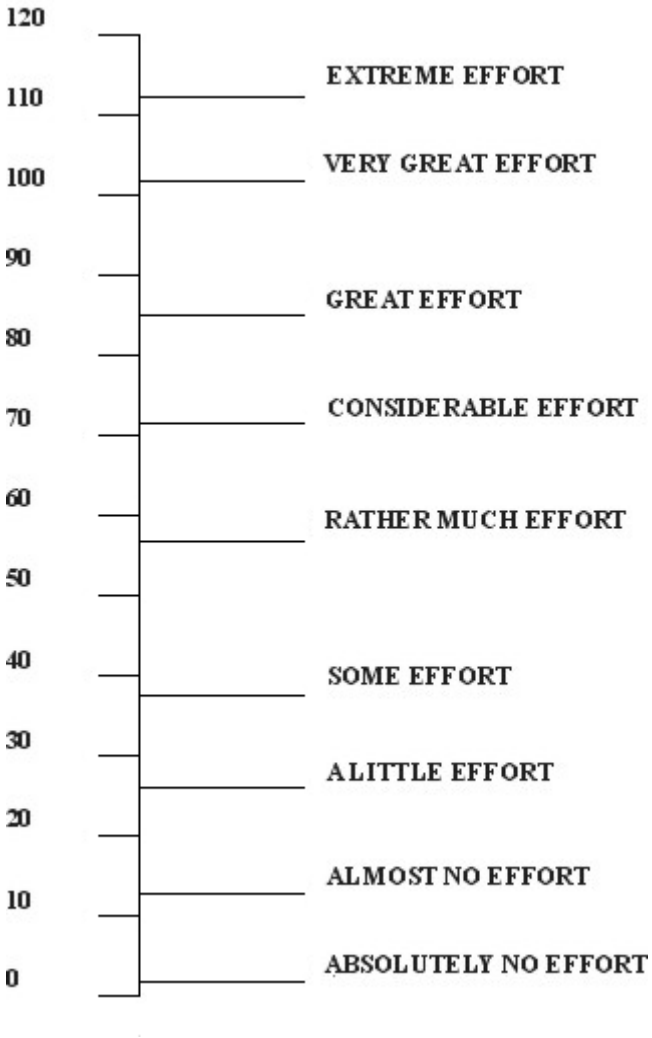


Figure 4.1 RSME modified for the experiment

An important aspect of the experiment are the camera images shown to the participants. The test vehicles from the ND project are equipped with 4 cameras: two of them pointed towards the driver, one from the front and one from the side; a front-view camera and a camera pointed towards the

navigation device. As a minimum, the front-view camera images will be displayed, with optionally a combination of the others added. It is decided to only use the video images from the front-view camera since this makes it easier for the participants to keep track of everything that happens during the situations. Having two displays will distribute the attention of the participants and may cause them to miss an important aspect in either one of the displays. The participants will also need less time when having to focus on only one display, resulting in a shorter video length and consequently allowing for a higher number of videos being rated. Another, more practical, advantage is that it allows the video fragments to be shared more easily, since any privacy concerns for the participants is sharply reduced.

In the determination of the length of the video fragments a number of considerations are made; if the videos are too short the participants do not have enough time to place themselves in the position of the driver and to make a judgment on what they have seen. However longer fragments make it much more difficult to control the variables, since they have to remain constant over a longer time period and no confounding variables may appear during the fragment. Furthermore, increasing the time of the experiment increases the time requirement of the experiment. At first, the videos used a time frame of only 10 seconds. However in a brief pilot study (5 participants) it was found that the 10 seconds was too short and video length was subsequently increased to 15 seconds.

In the creation of the video fragments from the naturalistic driving data, the primary concern was to limit the influence of confounding variables as much as possible. More specifically, the concern was to prevent the occurrence of any lane changing or curvature, which can be assumed to have a strong influence on workload. It was however difficult to account for effects such as environment complexity and the behaviour and presence of nearby vehicles, of which their influence is difficult to quantify. Screen captures from all video fragments can be found in Appendix A.

it was estimated that the average time that is needed to fill in the RSME is around 10 seconds. This period is likely the largest for the first few videos a participant has to judge, due to the inexperience with the scale. The participants are given an unlimited time to fill in the RSME, letting the next video start once the participant confirmed their selection. This results in the experiment having an approximate duration of 15 minutes for each participant.

The videos and RSME scale are integrated using a software application which was created for the sole purpose of this experiment. Figure 4.2 shows the interface of the program.

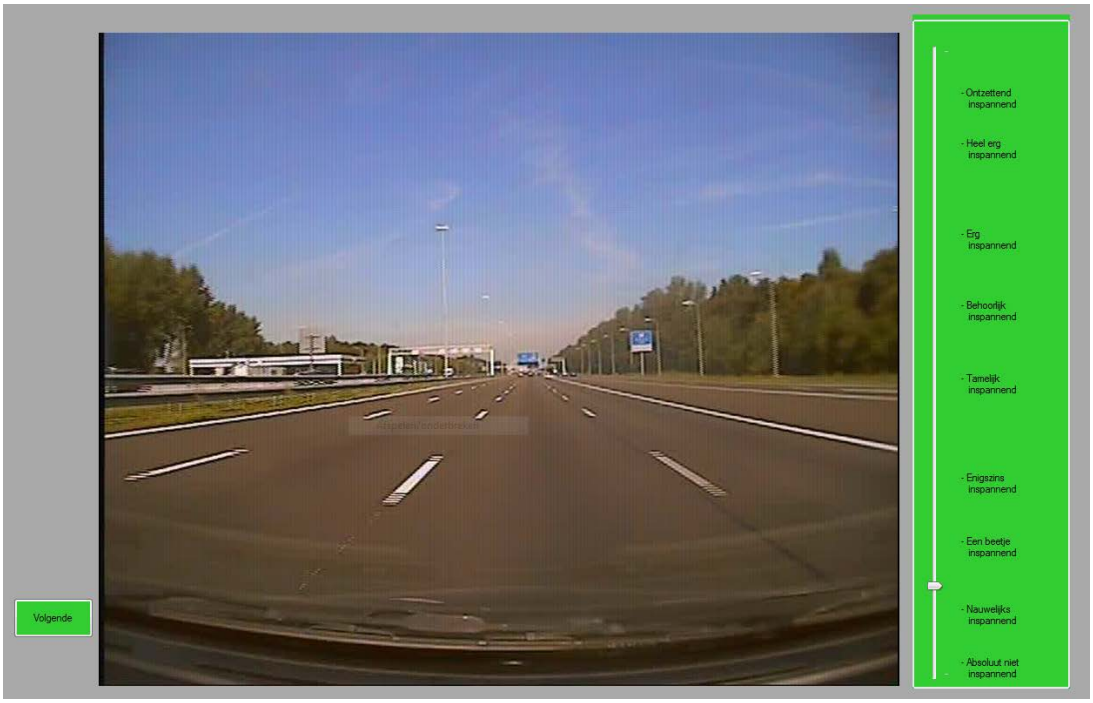


Figure 4.2 Experiment Software

The interface contains a start button, which starts the first video, a video screen which displays the video fragments and a slider which is used to fill in the RSME. After pressing start, the software chooses a video fragment at random and displays it for the duration of 15 seconds, after which the slider panel becomes available for use. After the slider is moved the 'Start' button, which turns into a 'next' button after first use, lights up and when pressed initiates the next random video fragment. This cycle will continue until all 32 videos are rated, after which a message is displayed thanking the participant for their participation.

4.3 Selection of variables for the experiment

In the previous chapter, a large number of different variables occurring in traffic, possibly having an effect on the driver's mental workload, were discussed. Since the goal in this research is to create a factorial design including all possible combinations of variables the total amount of video fragments that would be required is 2^14 = 16384 videos when including all variables in the research. With an estimated time of about 30 seconds per video this would require the participants to sit through 136 hours of footage, which they are unlikely to accept. For this reason it is decided to make a selection of variables of which combinations will appear in the experiment.

Only motor way situations are examined in the primary design The main reason for this is the similarity between different motor ways, allowing them to be more easily expressed as a combination of a limited number of variables. In urban situations especially the differences in environmental complexity can introduce a large amount of distraction. Environmental complexity is very difficult to quantify: how would for instance a roadside advertisement hold up against the presence of nearby buildings? For the motor way situations five variables were chosen: the traffic density, presence of heavy good vehicles, number of lanes, weather and the availability of natural light. The occurrence of difference driving maneuvers is not included, since they show case by case difference based on the behaviour of the both the driver and the surrounding traffic, resulting in a

reduction in control over the variable. In-vehicle technologies are also excluded from the experiment, since the participant is not operating the vehicle and cannot be aware when technologies such as ACC are active. While situations in which the driver is operating a navigation device or mobile phone are present in the ND video images, the dual task paradigm is not accurately conveyed to the participant whom only performs a single task (watching the video screen).

In a factorial design with 2 levels for each of the variables, the total number of video fragments is then 2^5 = 32, which at an estimated 30 seconds for each video resulting in a total time of 16 minutes. Due to practical and technical limitations it was however not possible to include the weather and natural lighting effect into the factorial design. The weather effect did not occur frequently enough to allow for every possible combination. Furthermore, rain often ranged in intensity. The front view camera showed very distorted images at night as a result of the head- and taillights of other traffic and roadside lighting, of which an example is shown in Figure 3.1. Furthermore it is doubtful that all possible combinations with the other variables could be found, considering the lower expected density during nighttime.

Figure 4.3 Nighttime naturalistic driving images

In the end, the variables traffic density, number of lanes and presence of heavy good vehicles remain. For each of these variables 2 levels are chosen, high or low density, 2 or 3 lanes and the presence of HGVs or no HGVs. This results in a total possibility of 2^3= 8 combinations. In order to increase the strength of the combinations and to test the assumption that the situations could be described with those variables alone, for each combination two video fragments are created. The average of the ratings of the two video fragments will then be used in the analysis.

Besides the combinations of motor way variables, an additional design is included containing mostly urban situations. These videos contain pairs of nearly identical situations, except for the inclusion of an additional experimental variable of which the main effect of their presence on mental workload is studied. The variables that are selected for this are the presence of Vulnerable Road Users (VRUs) and bad weather conditions. Apart from being able to study the main effect, it is interesting to notice if there is an effect of the order and spacing of the appearance of the pairs.

Since the experiment only contains a total of 32 videos, and the situations are extremely similar (except for the experimental variable), the participants notice that they have judged a video on the same location before. An additional advantage of including this design is that it may mask the purpose of the experiment when using only motor way situations. The participant may guess what the intended goal of the study is, by noticing the changing variables through the different combinations. This may affect the choices the participants make with regard to their perceived effort. The non-motorway situations are different to such an extent that the small differences between the motor way video fragments are less notable.

It is often assumed that the presence of VRUs creates an increase in the workload of the driver. Often VRU's are implemented in the surrounding in order to artificially create a similar yet higher workload (Cantin et al., 2009; De Craen et al., 2007; Edquist et al., 2012; Paxion et al., 2013). However little research has been performed on quantifying the influence of the presence of VRUs. While an extensive study on the effect of the presence of the VRUs cannot be performed with just these extra videos, it can be studied if a main effect can be observed. In the previously mentioned studies, the implementation of the VRU was done through a simulator or through the use of two different photographs in which the stage was set up, resulting in a perfect control over the present variables. In this research the situations are obtained through the use of similar surroundings in the naturalistic dataset. Different instances of daily occurring trips tend to have a very high resemblance, resulting in the possibility to find situations that are very similar yet different in the intended independent variable. While through extensive searching of the database, these settings can be found, the variables cannot be controlled to such an extend as would be possible in a simulator studies. This results in a slight loss of internal validity.

An example is given in figure 3.2, which displays a frame of two video fragments which are included in the experiment. As can be seen the surroundings are very similar and the effect of the studied independent variable is most noticeable (in this example the cyclist), however some other differences can be found as well. The distance to the lead vehicle is slightly larger in the right figure, and the sky is slightly more cloudy in the left figure. While these effect are expected to be of significantly less influence than the occurrence of the cyclist crossing the street, it should still be noted that this may have an effect on the results.

Figure 4.4 Similarity between extra videos - VRU

A second group of 8 videos is selected in which the weather is the only independent variable. Once again the situations are intended to resemble each other as closely as possible, with the exception of the weather. Figure 3.3 shows an example of this.

Figure 4.3 Similarity between extra videos - Weather

Appendix A.2 shows screen captures of all the extra videos, which shows the difference between each pair of video fragments.

4.4 Design

The main part of the experiment is a three way mixed factorial repeated measures design. Each combination of the within-subject variables: traffic density, number of lanes and presence of HGVs is twice represented in the experiment, after which subsequent analysis uses an average over the two scores. For each of the within-subject variables, two levels of their presence are distinguished. For the traffic density low density and high density are used, in which low density is determined to be traffic density where the speed of the vehicle is not affected and high density is near-capacity density. For the number of lanes, either two or three lanes are used and for the presence of HGVs either a single or multiple HGVs are shown over the course of the fragment, or none at all.

The second part of the experiment uses a mixed repeated measures design, in which two within-subject variables are used: weather and presence of VRUs. For both variables, eight fragments are included, which consist of (nearly) identical traffic situations in which the only difference is the presence of either rain or VRUs. A distinction between the type of VRU is not made, the videos contain either pedestrians or cyclists or a combination of the two.

The only between-subject factor on which a selection is made is the age of the participants, which is divided in three categories. These are young (18-25 years), middle-aged (30-50 years) and elderly (65+ years). Other between-subject factors subjects are their gender, driving experience in years, yearly kilometrage and the frequency with which they use the motor way. All participants recruited participated in both parts of the experiment.

While being part of a separate design, both parts of the experiment were not separately performed. Because of the use of a repeated measures design, a method of counterbalancing is applied in which the sequence of the fragments for each participant was randomized. This method includes both designs to increase the diversity and lead the subject's attention away from the purpose of the study. When separating both parts the similar nature of the individual videos belonging to each part may cause the participant to guess the purpose of the experiment.

4.5 Procedure

The experiments are all conducted using a 15 inch laptop set up in a room. The participant is seated behind the laptop, and is explained what activities are performed during the experiment. Before the participant is allowed to start the experiment, he or she is provided with an introduction. This introduction serves several purposes: firstly to explain what is expected of the participant during the experiment, secondly to familiarize the participant with the use of the RSME scale, and thirdly in order to give a reference for the use of the RSME for the first couple of videos.

Participants during the pilot indicated that especially for the first couple of videos they had trouble placing their RSME marker, having yet to build a frame of reference for themselves. After a small number of videos the participants are able to compare the new videos to the previous ones and are better able to secure a rating. For the building of the frame of reference, initially two aspects are of importance: the location of the place on the scale where the first rating is placed, and the intervals between different videos on the scale. The purpose of the introduction is to give an indication on these two points. In the introduction, two extremes are shown, which indicate the interval for which the RSME is intended to be filled in, and a different situation which falls somewhat in between the two.

The participant is informed that that they are about to see a series of video fragments, in which they are asked of them to place themselves into the position of the driver of the vehicle and make a judgment on the effort which is required to drive in that situation using the RSME scale on the right side of the screen.

An issue which was encountered during the pilot is the interpretation of the term effort (the Dutch word 'inspanning' was used in the experiment). A number of participants thought the term effort as synonymous with high effort, which resulted in a lot of situations which could be considered low effort to obtain a score of no effort at all. The affected ratings are easy to identify, since they have a tendency to fill in a score of 0 or 1 in some of the situations. Technically this would indicate that the

participant finds that he or she does not even have to be conscious in order to safely operate the vehicle, which is obviously misguided. A likely cause for this is the association of the Dutch word for effort with highly complex or even dangerous situations in traffic. Since in the experiment all situations require moderate effort at most to successfully negotiate, the participants rate these situations with a low score, anticipating more intensive situations at a later stage and thus reserving their higher scores for more complex situations which will never occur. They do not realize that the task of driving itself requires at least some effort, which may be partially caused by the method of examination, which does not let the participant drive an actual vehicle. In order to remedy this issue, the introduction was expanded for subsequent participations. This is done by explicitly mentioning that all the video fragments are depicting situations which occur in daily traffic and giving a greater elaboration on the definition of the term effort. Explaining that no effort means that in fact no effort is expended at all, i.e. the driver is expending as much effort as if they were sleeping, and relating it to the term attention, which has a more intensive-neutral association than the word effort, resulted in greater understanding in the use of the scale.

After this, the three introduction videos are shown. The first shows the vehicle driving on a near empty four lane road, with the purpose of displaying the minimal effort that the scale should be used for, and explaining that no effort means absolutely no effort and should not be possible during the driving task. Furthermore the use of the scale is shown, explaining that the slider could not be used until the video fragment has been fully played and showing that the slider does not have to place on either of the anchors, but could be placed along the full length of the scale. After showing an example rating, the ‘volgende’ (next) button is pressed, which initiates the next video and shows the self-paced nature of the experiment. The second introduction video shows the car driving through heavy snowfall, with a HGV moving in towards the driver. The third introduction video shows a three lane road with some other vehicles and a HGV, and a rating in between the two is given. Figure 4.3 shows images from these video fragments. After these videos the introduction is finished and the participant moves on to the actual experiment, which they perform independently.

Figure 4.5 Introduction videos

4.6 Chapter Summary

This chapter described the development of the method, which is used to measure workload in the experiment and attempts to answer the third research question: How do the results of the developed method compare to results found in workload studies performed in the past? In the experiment the participants watch a series of video fragments, each containing a combination of a number of variables of which an influence on mental workload can be assumed. Following each fragment, the participant is asked to indicate on a subjective rating scale their perceived effort as if they were driving the vehicle.

A total of 60 participants divided over 3 age categories: a young group (18-25 years), a middle-aged group (30-50 years) and an elderly group (65+); participated in the experiment. Besides their age, information was asked on their gender, driving experience in years, yearly kilometrage and the frequency of which they use the motor way. The participants are presented the video fragments through a software application created for the sole purpose of this experiment. This application presents the video fragments (15 seconds in length) one by one to the participant, allowing them to fill in a rating on the scale after each individual fragment. The subjective scale selected for this experiment is the Rating Scale Mental Effort (RSME; Zijlstra, 1993), because of its uni-dimensionality and sensitivity in high performance driving situations. The experiment is mostly self-paced, but does not allow the participant to fill in a rating until the full 15 seconds of the video are played out, and does not allow the start of the next video until the slider of the RSME is moved. All participants rate all video fragments, which are presented to the participants in a random order as a means of counterbalancing.

The experiment features two designs. The first design is a factorial design of the three variables: traffic density, number of lanes and the presence of HGVs. For these fragments only motor way situations are selected and each combination is included twice. The reason to only include motor way situations is because they can easily expressed as a combination of a limited amount of variables and suffer little from environmental complexity, which is difficult to quantify due to the wide range of possible sources of complexity. The second design contains pairs of near-identical situations of which the only difference is the addition of either a single or multiple vulnerable road users (VRUs) or poor weather (rain) to one of the pairs. This allows for the study on the effect of the presence of these variables on mental workload, as well as a secondary purpose which is the masking of the first design. When the experiment would contain solely slight variations in the motor way situations, the participant may catch on to the purpose of the experiment which could influence their decisions.

Before the participant is allowed to start the experiment, a short introduction was provided to them. This introduction services several purposes: it explains the procedure of the experiment, familiarizes the participant with the use of the RSME scale, and gives a reference for the use of the RSME for the first couple of videos. Participants during the pilot indicated that especially for the first couple of videos they have trouble placing their RSME marker, having yet to build a frame of reference for themselves. After a small number of videos the participants are able to compare the new videos to the previous ones and are better able to secure a rating. This is done by showing the participant three example situations created for the purpose of the introduction; an expectantly low effort situation, a high effort situations and a situation which falls somewhere in between these two. For each of these examples an indication is given on the RSME scale where the situations could be

scored. After the introduction is provided the participants start on the actual experiment, which they perform independently.

5 Results

In this chapter, the results of the experiment as described in Chapter 4 are presented and discussed. First an initial look at the data using descriptive statistics is taken. Next, a mixed repeated measures ANOVA (rANOVA) is used to evaluate the influence of the separate and interaction effects of the variables. Following this, the results from the extra videos, relating to the influence of the presence of VRUs and the effect of rain on mental effort are analyzed. In the end some additional results are shown which include the variables: gender, .

5.1 Descriptive analysis

Before any tests are applied, a descriptive analysis is performed in which the participants and their characteristics are examined and the results of the individual video fragments are shown. Furthermore is it tested whether the assumptions which are required for the performance of rANOVA are met.

5.1.1 Participants

The general information on the different age groups which was gathered during the experiment is shown in table 5.1.

Table 5.1 - Characteristics of the participants

Age group	Mean age (years)	Mean driving experience (years)	Gender distribution (male/female)	Mean kilometrage (km/year)
Young (18-25)	22.8	4.2	10/10	3793
Middle (30-50)	43.0	23.2	13/7	13267
Elderly (65+)	71.7	49.6	13/7	11382

As can be observed from the table, the concentration of subjects is slightly centered towards the higher half of the 18-25 and 30-50 range. In the setup of the experiment it has been attempted to keep the gender distribution close to even, however recruitment was not specifically focused on this. The table shows a big difference in the kilometrage among the different groups, which results in that any analysis on the effect of kilometrage needs to be controlled for age covariance. Information on kilometrage was not a mandatory question, since to a lot of people this number is not known to them, resulting in unavailability of this data for a portion of the subjects.

In order to test whether the sample is an accurate representation of the populace the characteristics are compared to Dutch traffic statistics gathered by CBS (Centraal Bureau voor de Statistiek). On average a 15-25 year old person drives 2978km, a 25-45 year old drives 9114km and a 65+ old drives 3146km yearly. For the younger and middle aged group this is comparable to the findings in this

research, however for the elderly group this number is much higher in the sample used in this research. These differences are probably because in order to be able to be eligible for this research the participant needs to regularly drive. It is likely that a large number of elderly seldom driver, bringing down the average for the whole category, which explains the big difference for this category in particular. Furthermore it is possible that elderly who are still driving regularly show a greater interest in participating in traffic safety research. The 13/7 gender distribution which is used in the middle and elderly age category is not that far from the gender distribution of the populace, which is on average 11910km/year for male- and 8778km/year for female drivers, which comes down to a 11.5/8.5 distribution.

In table 5.2 the mean score, SD score and mean range over which the RSME is displayed. This is included, since a notable difference was found between the groups. The younger group on average scored significantly higher than the middle aged group (p=0.0008) and also uses a larger range of the RSME than the middle aged group (p=0.0006) and elderly group (p=0.04). The comparison between the middle aged group and elderly group showed no significant differences.

Table 5.2 - Mean scores of Age Groups

Age	Mean score	SD score	Mean minimum score	Mean maximum score	Mean range
Young (18-25)	40.0	17.2	14.4	77.9	63.5
Middle (30-50)	26.9	11.8	9.3	54.1	44.8
Elderly (65+)	31.7	13.4	9.3	60.2	50.9

5.1.2 Video fragments

Table 5.3 shows the mean scores and SD scores for each of the videos, plus the average of the two identical videos which will from here on be used in the analysis. The abbreviations: H, L, 2, 3, N and Y stand for the density (High or Low), number of lanes (2 or 3) and presence of HGV's (Yes or No). For a look at the complete dataset, the scores given by every participant are shown in Appendix E.

Table 5.3 - Mean scores Video Fragments

Situation	Video	Mean Score	SD Score	Average
2-H-N	2hn1	33.7	19.4	35.7
	2hn2	37.5	19.2	
2-H-Y	2hy1	34.8	16.3	38.1
	2hy2	41.3	20.7	
2-L-N	2ln1	14.9	10.4	17.1
	2ln2	19.1	14.0	
2-L-Y	2ly1	25.7	14.5	24.7
	2ly2	23.6	14.6	
3-H-N	3hn1	36.8	18.9	34.4
	3hn2	31.9	17.6	
3-H-Y	3hy1	39.8	22.3	38.4

	3hy2	37.0	20.7	
3-L-N	3ln1	17.8	11.7	15.9
	3ln2	14.0	9.4	
3-L-Y	3ly1	24.0	15.1	25.1
	3ln2	26.0	15.7	

It can be observed from the data that substantial differences are present between high density situations and their low density counterparts. The inclusion of HGVs seems to have the most sizeable effect in the low density situations and the influence of the number of lanes seems limited.

A number of paired Student's t-test show that four significant difference were found between the supposedly identical situations. The implications of this are further discussed on paragraph 6.1. Pearson's-r tests between the combination show correlations ranging from r=0.542 to r=0.812. The full results of these tests can be found in appendix C.

In order to be able to use a rANOVA, some assumptions have to be satisfied. Beside the standard assumptions that the data should be measured at at least an interval level, and the data from different participants should be independent, the rANOVA also requires sphericity, which requires equal variance in the different levels of each variable. Since the maximum amount of levels is two, sphericity is automatically assumed. Furthermore normality of the data is required to a certain extent. This is tested by performing a Kolmogorov-Smirnoff test, including a factor for the different age groups, since they have been found to use the scale over different ranges. The normality tests show no significant results for most of the combinations of age groups and traffic situation distributions. However there are still some combinations which are significantly different from the normal distribution. A log 10 transformation was performed on the data but showed no definitive improvement. It is decided to still use the rANOVA, since the ANOVA is rather robust for deviations from normality (Schmider et al., 2010) and there is no non-parametric counterpart of a factorial ANOVA (Field, 2009). Both the results of the initial normality tests and the log 10 transformation can be found in appendix B.

5.2 Repeated measures ANOVA

With the scores from the research known, it is not possible to analyze the results from the experiment using a rANOVA. The main effect of within subject factors: density, number of lanes and the presence of HGVs and between subject factors: age, gender and kilometers driven per year as well as their interaction effect are examined. The advantage of using a rANOVA is that the variance as a result of different participant's tendency to place their ratings over a different range is taken into account. All pairwise comparisons are performed using the Bonferroni adjustment.

Besides finding the statistical significance of the change, it is also interesting how large the influence is of the individual factors and how they relate to each other. The results from the RSME scores are expressed in mm, which in itself does not say a lot about the actual workload of the subject, especially considering people have a tendency to fill in the RSME over different ranges as a result of different interpretations rather than an actual different in perceived mental effort. For this reason it is better to look at standardized effect sizes. In table 5.4 the standardized effect sizes (partial eta-square) are shown, with an asterisk indicating a significance level of p=0.05 or lower and a double asterisk indicating p=0.01 or lower. According to Kirk (1996) an effect size is small when it is has a

value of 0.01, medium at 0.06 and large at 0.14. The full results including F values and degrees of freedom can be found in appendix H.

Table 5.4 Results rANOVA (values are standardized effect sizes (η^2))

	Main Effect	Density	Lanes	HGVs	Age
Density	0.749**	x	0.001	0.162**	0.121*
Lanes	0.014	0.001	x	0.035	0.142*
HGVs	0.465**	0.162**	0.035	x	0.065
Age	0.163*	0.111*	0.142*	0.065	x

5.2.1 Main Effects

As can be observed from Table 5.4, three main effect have been found: Density, presence of HGVs and age. The results shown in the table confirm the observations which were made in paragraph 5.1.2. In figure 5.1 the effects of the within subject variables are graphically displayed.

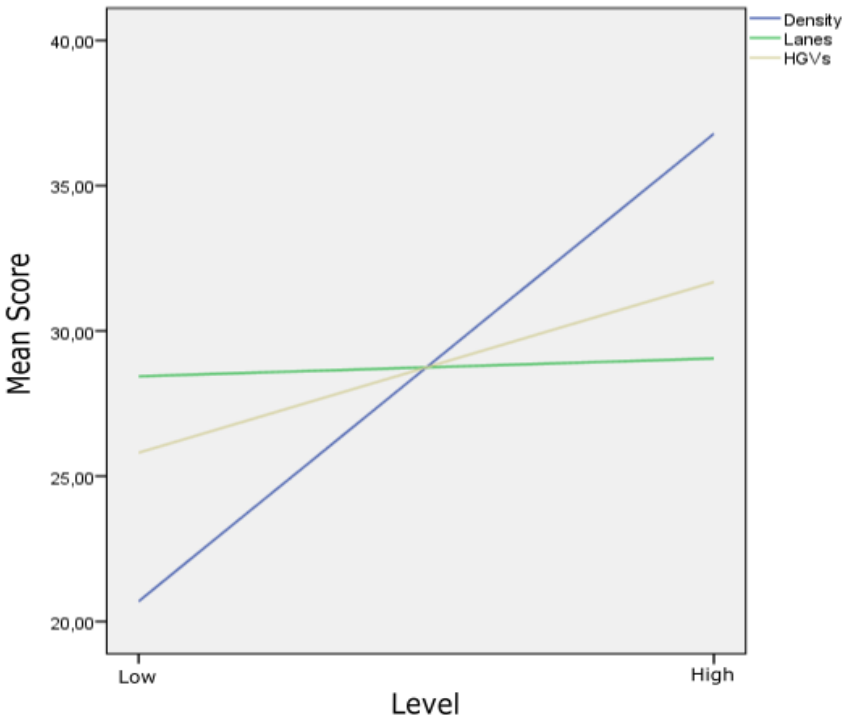


Figure 5.1 Main effects within-subject variables

In the figure, low level variables means 3 lanes, with high level meaning 2 lanes. For HGVs, low level means no HGVs present and high levels means HGVs present. As shown in the figure, the effect of density is much larger than the effect of the other two. A significant main effect was also found for the between subject variable age. In figure 5.2 this effect is shown.

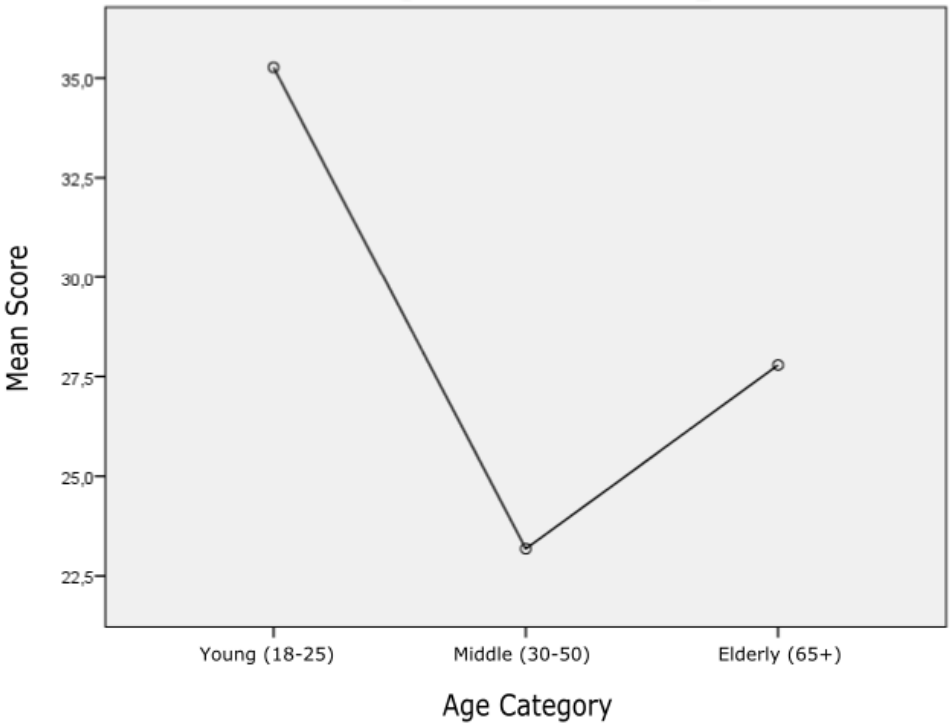


Figure 5.2 Main effect Age Category

From the figure it shows that the youngest group indicates the highest mean subjective effort, with the elderly group a reduced amount and the middle-aged group even lower. Pairwise comparison however shows a significant difference between group 1 and 2 only.

5.2.2 Interaction effects

Besides the main effects, some interaction effects were also found. A significant interaction effect was found between the density variable and HGV variable. Figure 5.3 shows this effect.

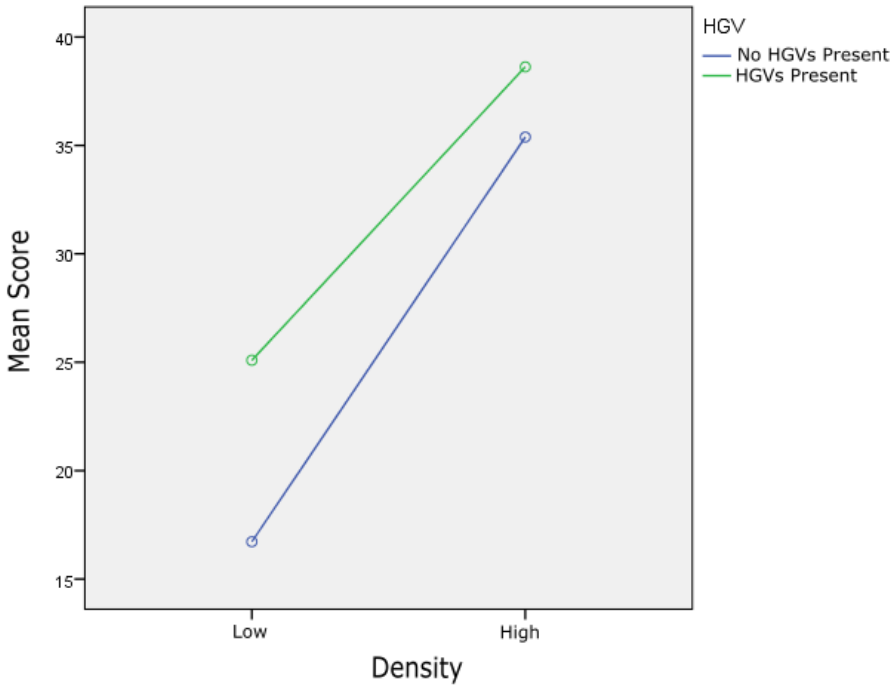


Figure 5.3 Interaction effect Density x HGV

It can be seen that at a low density, the difference in mean scores between both conditions is much larger than it is at high density.

Furthermore an interaction effect was found between the Density and the Age category. This is depicted in Figure 5.4.

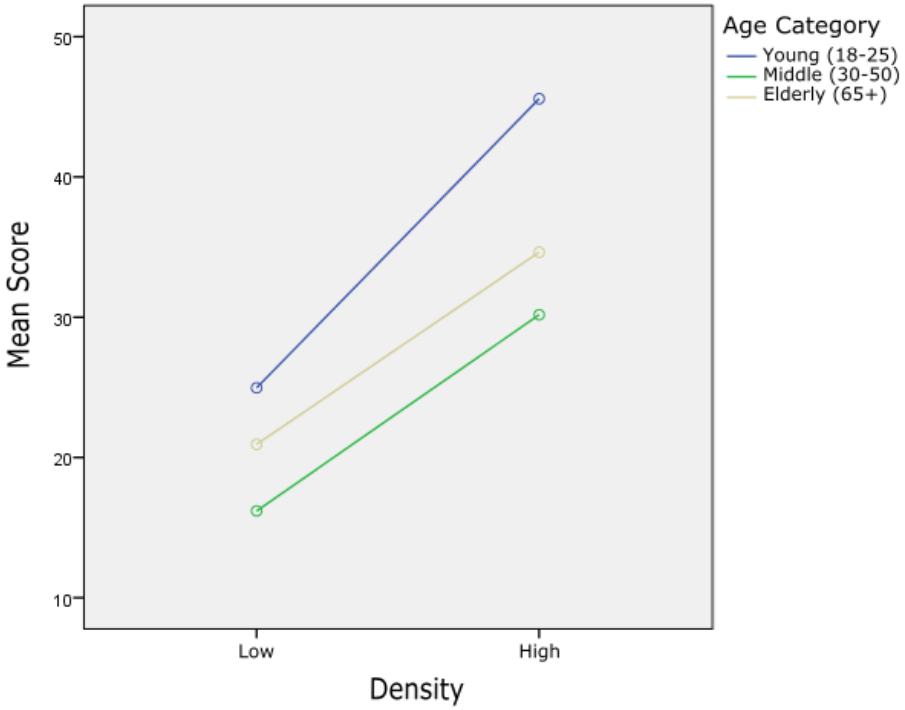


Figure 5.4 Interaction effect Density x Age Category

As can be observed from the figure, for the middle aged and elderly subjects, the means increase by about an equal amount for increased density, however for the younger subjects this increase is much larger.

Figure 5.5 shows a plot of the interaction effect between lanes and age of the participants.

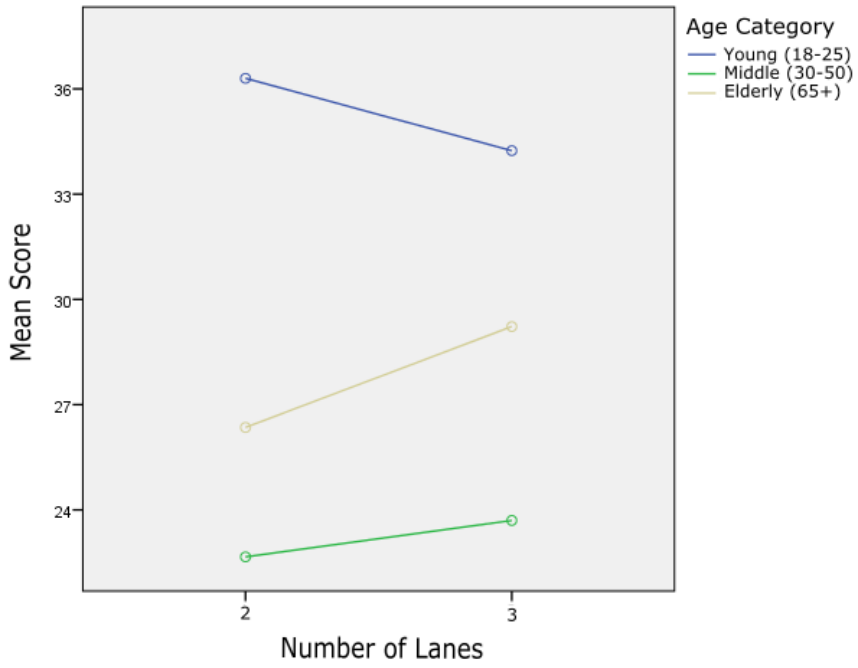


Figure 5.5 Interaction effect Lanes x Age Category

As can be observed for the younger group, the subjective effort actually decreases with a decrease in the number of lanes, as opposed to the other two groups where a lower amount of lanes is found to require more effort. In Figure 5.6 this effect is shown for different density levels.

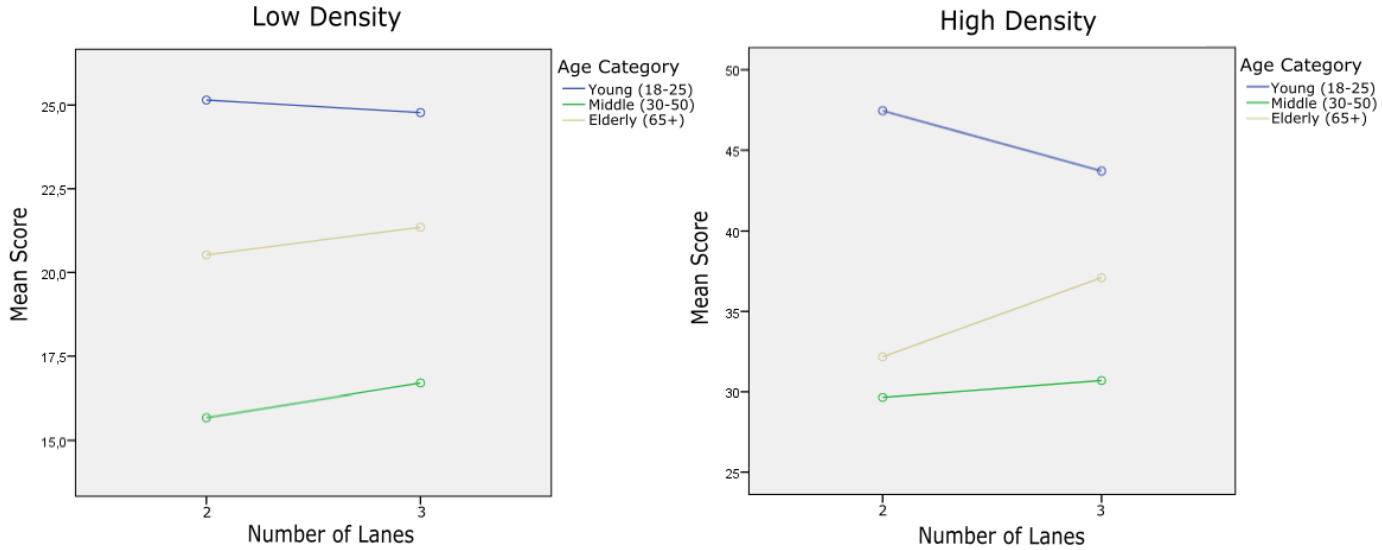


Figure 5.6 Interaction effect Lanes x Age Category x Density

The figures show that the effect which was shown in Figure 5.5 is mostly apparent at higher density. For lower density the subjective effort reported by the younger group is not that affected by the number of lanes, however at higher density it increases. While the Age x Lanes x Density interaction

effect was not statistically significant (p=0.131), an effect size of η^2=0.069 is found, which indicates a medium effect size.

5.3 Extra videos

Besides serving as a method to direct attention away from the motor way situations, the extra videos are also used to find a main effect of presence of VRUs and weather conditions on driver's mental workload. The results of VRU and Weather effects are again analyzed using a repeated measures ANOVA, however in these instances with a single factor and 4 measures. For both the VRUs and Weather effects a significant main effect was found.

Besides the main effect, the interaction effect with age is also studied. Table 5.5.5 shows the results of the Repeated Measures ANOVAs. Again the standardized effect sizes are shown (partial eta-square) with an asterisk indicating a 0.05 significance interval and double asterisk indicating a 0.01 significance interval.

Table 5.5.5 Effects Presence of VRU and Weather Effects (values are standardized effect sizes (η^2))

	Main Effect	Interaction Effect with age
VRU	0.732**	0.216**
Weather	0.680**	0.065

As can be observed from the table, a main effect of both VRU and Weather effects was found. No interaction effect from age with weather was found however, and coincidentally when looking at the main effect of age using only the data from the extra videos alone, only a significant effect was found for the VRU data.

Figure 5.7 shows the interaction effects for the four measures of the VRU x Age category.

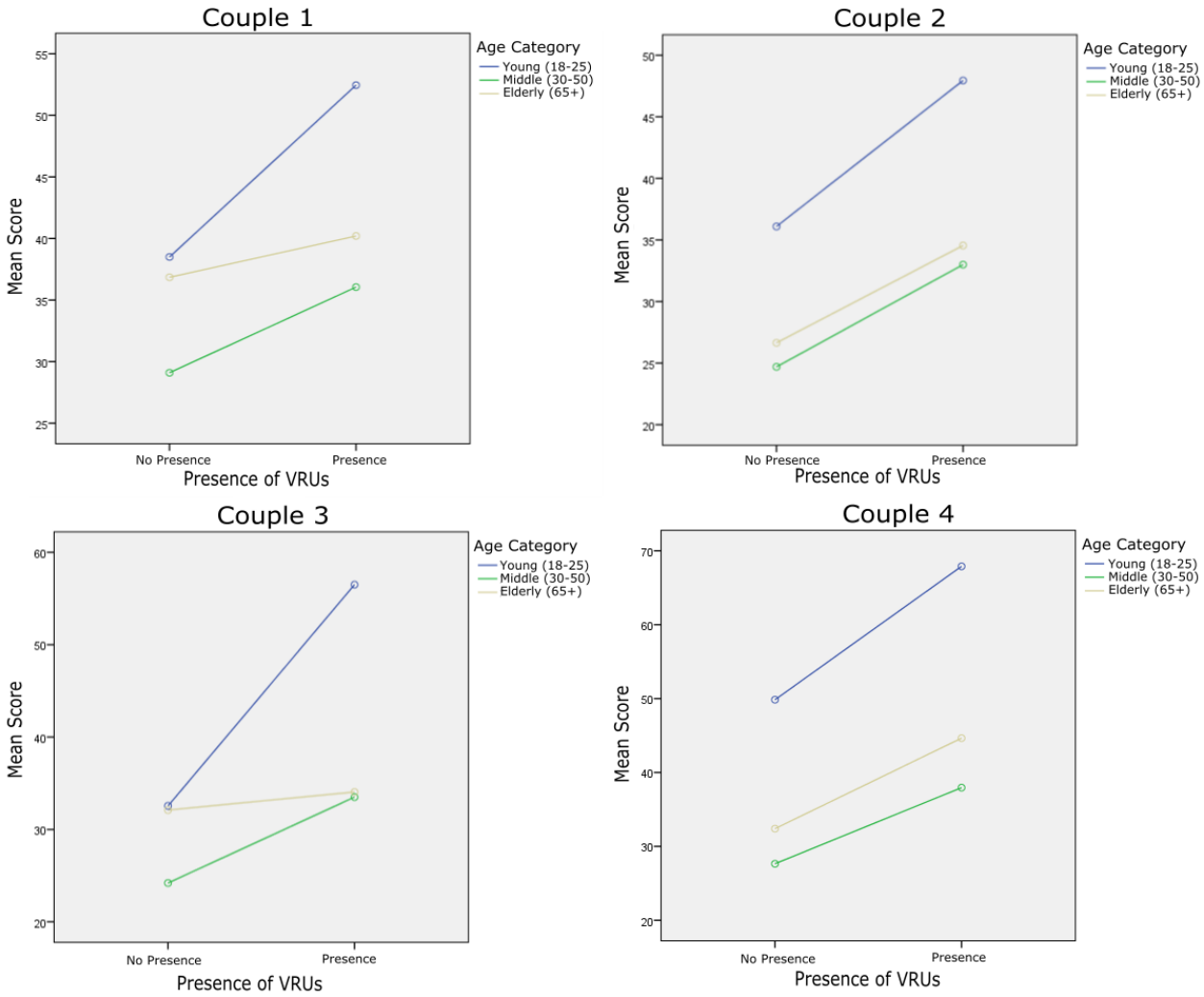


Figure 5.7 Interaction Effect VRU x Age Category

The graphs are separated for the four different couples. As can be observed from the figures, the younger group tends to have a relatively stronger reaction to the presence of VRUs, with the elderly group reacting differently from situations to situations.

5.4 Post hoc analysis

In this paragraph some additional analysis is performed aimed at finding correlations using data which was not originally part of research.

An additional attribute of the videos which is yet to be used is the speed of the vehicle driven in the video fragments. This could be used as an additional independent variable. With the average speed known for each for the motor way situations, a Pearson's-r test could test whether a correlation between the speed and average rating of the videos can be found. Since information on speed is not known for all the situations, three situations are left out, namely a 2LY and both 3LY situations. A negative significant correlation was found between the speed and average scores (r=-0.719**). Since visual information enters the perception of the driver with a speed which is a function of the speed of the vehicle, it is expected that higher speeds should result in a higher workload. It is however

likely that, since the correlation was found to be negative, the values are resulted from the relation with the traffic density. Traffic flow theory dictates that an increase in density is accompanied by a decrease in speed. Since the high density fragments were accompanied by the highest reported subjective effort and the lowest speeds, it is likely that density is the main cause for this. A second Pearson's-r test is performed using the standard deviation of the speed, however this shows no significant correlation with either of the groups. Additional analysis on the effect of speed as a possible confounding variable is performed in paragraph 7.1.

Characteristics of the participants which are known are gender, driving experience in years and km/year driven, and the number of times the person uses the motor way each month. Here, especially km/years driven and gender are interesting to study; driving experience in years shows a very strong correlation with age (since most participants received their driver's license at age 18 or 19) and the number of times per month participants drive on the motor way is often misinterpreted, as people have an tendency to forget to include the retour trip.

The Repeated Measures ANOVA on the motor way data is once again performed, however this time including the gender as a between subject variable. The rANOVA did not return a significant main effect. However a significant interaction effect was found between age and gender ($p=0.015$, $\eta^2=0.145$). This effect is displayed in figure 5.8. As can be observed from the graph, the self-reported mean rating actually decreases when comparing the middle-aged to the elderly group, while for the male group the means increase.

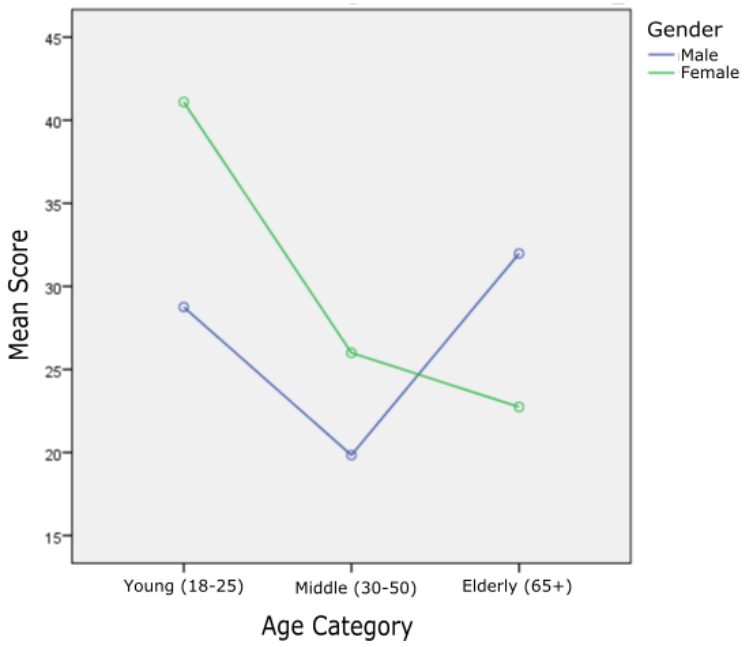


Figure 5.8 Interaction effect Gender x Age

Before a similar test using the kilometrage is performed, first a categorization must be made between experienced and inexperienced drivers. In Patten et al., 2006, the participants were categorized as experienced drivers with an annual kilometrage of 15,000 or above and inexperienced at 15,000 or below. A first test performed using the same classification and including a covariate for age category, revealed no significant effect for either the main effect or interaction effect. This

contradicts with the results found in Patten et al., (2006), in which a significant difference between workload for experienced and inexperienced drivers was found. However their experienced drivers were all professional drivers whom to my knowledge do not appear in this research.

5.5 Chapter Summary

In this chapter the results of the experiment were displayed. First, some descriptive analyses were performed. Striking is the difference in ratings given by the different age groups. The youngest group provided, on average, a significantly higher rating to the video fragments, followed by the elderly and the middle aged group respectively. Furthermore it was noticed that there the younger group has on average a lower kilometrage, which means that the analysis needs to be adjusted for covariance.

The results of the experiment were analyzed using a Repeated Measures ANOVA (rANOVA). Significant main effects ($p<0.05$) in the motor way situations were found for: traffic density, presence of HGVs and Age of the participant. Participants found that the high density situations required much more effort than the low density situations. The presence of HGVS was also found to have a statistical significant influence; however the number of lanes did not result in significant results. In the experiment the young group indicated to have the highest mean workloads, followed by the elderly group, and the middle-aged group indicated to require the least amount of effort in the explored situations. This effect is explainable; younger drivers lack the experiences and the development of automatic action patterns present in the other groups, while older drivers have a reduced processing capacity as a result of the aging process. However in studies using subjective workload measures it is often prevalent that the younger group overestimates their driving skill, which was not found in this research. Interaction effects were found for Density x HGVs, Age x Lanes and Age x Density. The effect of HGVs was greatest in low density situations. The younger group found 3 lane situations to be more effortful than 2 lane situations at high density, while the other two groups found the 2 lane situations to be more effortful. Furthermore the younger group showed a higher relative increase in mean workload as a result of increased density than the other two groups did.

rANOVA on the secondary videos showed a main effect of the presence of VRUs and adverse weather, with an interaction effect found for VRU x Age; again having the younger group indicate a relatively higher increase in mean workload in the more difficult situation. No interaction effects were found containing the weather variable. In a post-hoc analysis a significant correlation was found between the speed driven in the video fragments, and the average subjective ratings. This is however most likely caused by the relation between speed and traffic density, which was found to be the strongest prediction for workload. A rAnova performed including the gender showed an interaction effect between age and gender, but no main effect. No significant effects as a result of difference in kilometrage were found.

6 Evaluation of the measurement method

Having applied a relatively untested research method, which has yet to be validated for practical use, it seems appropriate to pay attention to the evaluation of the method. While no validation study has been performed, some aspects of the results can still be evaluated. The evaluation made in this chapter is based on remarks made by the participants and certain patterns found in their ratings. The first part of the chapter focuses on the validity of the method. The different types of validity are explained and for each type some concerns are provided. Afterwards some artefacts which may have been constructed as a result of the experimental setup described in chapter 4 are named and the significance of their introduced bias is tested.

In this chapter an evaluation is provided on both the method which was devised for this experiment, as well as the use of the method in this particular study. Mistakes that are made in the particular research should not necessarily reflect negatively on the experiment method, but are often better attributed to a lack of experience or insight of the user.

6.1 Validity

Validity determines to what extend a concept, conclusion or measurement accurately conforms to the real world. Three types of validity can be distinguished: test validity, experimental validity and diagnostic validity; of which each of the types have certain aspects that determine their respective validity.

Test validity is the extent to which a measurement measures what it is intended to measure. This should not be confused with the related term reliability, which is the degree to which the results of a measure are consistent over multiple repetitions. Three aspects that are part of test validity are: construct validity, content validity and criterion validity. Construct validity is the extent to which a test measures what is it designed to measure. For this particular study it determines to what extent the method will actually be measuring mental workload. Content validity determines to what degree the method measures every aspect of the subject that is examined. For criterion validity the test is compared to other methods which in the past have been proven to be valid and is a combination of the predictive validity and concurrent validity. The difference between these two is that for concurrent validity the two measures are taken at the same time, while for predictive validity one measure is made at an earlier time and servers to predict a later measure.

Experimental validity determines whether it is possible to draw valid scientific conclusions based on the results that are obtained in the experiment. There are three aspects of experimental validity: statistical conclusion validity, internal validity and external validity. Statistical conclusion validity is determined by the use of statistical techniques when drawing conclusions. It is important to use a correct sample size and use appropriate statistical tests in order to be able to draw conclusion based on the experiment results. Internal validity determines whether based on the experimental setup conclusions on causal relationships can be drawn. An important aspect is the control over the variables present in the experiment, in order to isolate the influence of an independent variable on a dependent variable. External validity concerns to what degree the results of an experiment are true for other cases. An example of this is whether the sample that is chosen is a good representation of the general population.

In the third type of validity, diagnostic validity, the validity of a diagnosis is assessed. Due to the nature of diagnosticity, certain symptoms of a condition points towards multiple conditions, the methods seen to be sensitive enough to detect a problem, but specific enough to not respond to other things. This type of validity is however rather specific to clinical fields, and will therefore not be further specified in this report.

In this particular research there are uncertainties on some of these validity types. In applying this method, construct validity seems like the most important aspect. While we ask the participant to indicate their perceived effort as if they were in the driver's place, there is no guarantee that the participant is actually able to reason from the driver's point of view, resulting in that what is indicated might not be the mental effort at all. For this it is especially important that it is clear for the participant to know what is asked of them. All participants must have the exact same interpretation of what it is that they are commenting on, and that interpretation must obviously be in line with what we want to know of them. It is therefore important to clearly instruct the participants on the working of the RSME, and their role in the experiment.

There are two primary reasons which could cause a low degree of test validity. Firstly, only a self-report measurement is applied, while it is often encouraged to apply a second method (preferably performance based measures) in conjunction with self-report measures (De Waard, 1996; Wu and Liu., 2007). A dissociation between self-reported workload and performance can be found, especially in dual-task situations, since the participant often fails to account for the additional workload associated with multitasking (Horrey et al., 2009). This dissociation is not testable in a design in which only the self-reported workload is measured. Secondly, the participant is not the person whom is actually driving the vehicle, resulting that the results may be different from what a person who is driving at the time would indicate. An approximation on the validity can therefore only be approached by comparing the effects which are found in prior studies which employ methods of obtaining workload which contain better validity. In Chapter 3 a review of the literary was performed aimed at determining the effects of the variables studied in this experiment. Influences of density, HGVs and age on mental workload were found, which is in line with the results obtained through the method described in this thesis. No information was found on the effect of the number of lanes, which neither confirms nor denies the results found here. The interaction effect which was found between density and HGVs contradicts with what was hinted at in the literature study. Here a larger effect of the presence of HGVs was found at lower workload levels, whereas in the literature study a larger effect was found for the higher workload levels. This could however be attributed to the fact that the maneuvers performed in De Waard et al., 2008 and 2009 require a much higher effort than the one performed in this experiment (driving straight ahead), and thus the situations are not entirely comparable.

By using only a subjective measurement method there is a possibility that the results are influenced by bias. There is a tendency for some people to overestimate their driving abilities, particularly the elderly (De Waard et al., 2009). This may especially be prevalent with this method, where a lack of driving skill from the participant is not necessarily correlated to a decrease in the driver's performance. Opting to let the participant watch video fragments of traffic situations, rather than having them drive through the situations themselves also has an effect on the validity. When making an indication of the workload which was expended in the experiment, subjects relate to their

performance and effort while performing task when using the scale. Since no actual performance is made it may cause a bias in the results.

Judging from the arguments given above, obtaining any definite form of test validity using only the results from the experiment and no additional measurements being taken seems unlikely. Therefore in the remainder of this chapter the focus is placed on the experimental validity, which is to a certain extent possible to be determined using the results of the experiments and the information that is known of the participant.

Due to the lack of the distribution of the scores on the RSME related to some video fragments, a reduction in statistical conclusion validity can be assumed. Failure of the assumption of normality when using an ANOVA could result in the occurrence of a type II error, which represents a false negative. The ANOVA has proven to be rather robust to departures from normality however (Schmider et al., 2010), so it is unsure whether the results are affected by this.

Internal validity, and specifically the control over the variables, plays likely the most important role in the experiment set up. Because of the nature of naturalistic driving, there is no control over the traffic situations in which the vehicle drives and the variables that occur during these situations. This is for a part counteracted by the sheer size of the available dataset, resulting in that every possible combination of the variables occurs multiple times. It is not entirely certain if the variables that occur could be isolated to an extent that the influence of confounding variables is negligible. In the collection of the video fragment, an attempt was made to have the situations contain as little confounding variables as possible. Here, was focused mostly on the maneuvers which the driver partakes during driving, such as overtaking or changing lanes, and the prevention of finding HGVs on the other carriageways in no HGV situations. Due to a combination of a limited time period and an extensive catalogue of trips to choose from the naturalistic driving dataset (which contains hundreds of hours of footage), not every motor way minute has been extensively checked in search of the optimal footage. Compromises have been made mostly on the presence of roadside distractions, especially in scarcely occurring situations. A quick glance at the screenshots displayed in Appendix A already show some differences between the two supposedly identical situations.

The decision was made to include two versions of each combination of variables for the motor way situations, with the reason that an average over two fragments would result in a more valid score. Since the situations contain the same variables, it is desirable that the mean scores received from the participants are very similar. However for some of the situations, statistical significant differences were found between the two identical situations. This is probably the result of differences which are present in the identical situations as a result of poor control over the variables. One way to examine the influence of the variable control in this research is to use different single video fragment results instead of using the averages which were used in the rANOVA. By comparing the effects which are found through this method to the results shown in Chapter 5, it can be found to which end the results shown in Chapter 5 are a result of poor variable control.

Not each possible combination of videos is used in the rANOVAs, since that would result in a total of 2^8 = 256 combinations, however a selection is made based on the mean scores, which can be found in paragraph 5.1.1. A combination of the all the highest scores received per couple, the lower scores, the scores which result in the smallest amount of distance between the means and the results which

result in the greatest total distance between the means. The results of the rANOVAs can be found in Appendix D.

The result of the rANOVAs show different significance levels for the interaction effect for different combinations, however the main effect of density, HGVs and age the results are still significant for each combination. The conclusion which can be drawn from this test is that the differences in the individual fragments are sizeable enough to make a difference in the results when taken separately. Improvements on this account could be implemented in two ways: setting up the video fragments rather than depending on already made footage significantly increases the control over the situations or increasing the total amount of videos for each situations (currently only 2) and taking an average over a greater amount of video fragments should somewhat filter out the unwanted influences which are introduced by the individual videos.

For the external validity, the sample which is used should make an accurate representation of the overall population. This can be done by trying to find a sample of people that are representative of the total population in relation to their personal characteristics. While a selection criterion was made on the age of the participants, no selection was made on other personal characteristics such as driving experience, gender, social economic status or marital status. While information on driving experience and gender is known, no selection on it was made. A bias is likely found towards a higher level of education, since a lot of participants were recruited among employees and students at the VU university in Amsterdam. Furthermore recruitment was centered around the cities The Hague and Amsterdam, resulting in a high prevalence of participants whom are likely familiar driving in city surroundings. The possible bias introduced by this is however not testable without performing a similar study using a sample with different characteristics.

In Chapter 5 a correlation was found between the speed and the workload indicated by the subjects. While this is likely the result of the direct relation between speed and density, with density being the primary predictor of high workload found in this experiment, it is still possible to test whether the difference in speed could act as a confounding variable. The speed of the vehicle is an important factor in workload; when a person drives a constant flow of information is entering the persons processing system, of which the size of this flow is determined as a factor of the speed which is driven. For this reason people tend to generally reduce their speed when driving in a high workload environment. The effect of speed is tested by comparing the identical situations to each other and relating their difference to the difference in speed, for the situations in which the speed is known. A Pearson's-r test performed between the absolute difference of the speed and the scores of the fragments resulted in no significant results (r^2=0.337, p=0.227) however the sample size used in this test is rather small, after filtering out the situations for which no speed was known only 6 combinations remained. Based on these results no conclusion can be drawn on the effect of speed as a confounding variable in the experiment.

6.2 Artefacts in the measurement method

Since it is the first time this measurement method is used, it is important to examine possible causes of bias which are introduced through the attributes of the method. In this paragraph some possible artefacts are studied, which are a result of the subjects inexperience with the scale and the term mental effort, and the negative effects which may be introduced through the use of a repeated measures design.

6.2.1 Influence of the first video

When indicating their effort on the rating scale, scores of subsequent videos are often related to previous videos, scoring slightly higher scores when they are found to be more effortful than the last video, or slightly lower when they were found to require less effort. These intervals are generally quite small, meaning that the placement of the initial video could have a large effect on the placement of all subsequent ratings. It is therefore possible that the participants which started off with a video fragments that requires low effort (such as low density, 3 lane roads with no HGVs) have an overall lower score. This is tested by comparing the difficulty of the first video for each participant to the average score they give over all video fragments. Determining the overall difficulty of a video fragment is done by taking the mean score the video has received over all participants. A linear regression performed comparing the indexed difficulty for each participant shows no significant effect for either the means scores (r^2=0.002, p=0.734) or the standard deviation of the scores (r^2=0.0004, p=0.888). This means that it is unlikely that there is bias resulting from the first video shown to the participants.

6.2.2 Video fragment couplings

The term coupling in this report is used for both the supposedly identical situations in the motor way fragments as well as the extra video couples. For these videos their similarity it is often noted, especially for the extra video fragments, and sometimes the participant even remarks that a video is played for the second time and wonder whether it is because of an error in the program. Therefore it seems interesting to test whether there is any bias resulting from the sequence in which these similar fragments are presented to the subject. Two properties are tested, the influence of the distance between the two similar videos and the order in which they appear.

When two videos belonging to the same coupling follow each other in close proximity, the participant may remember what they had entered in the first video which may influence the score in the second video. Participants often remark that they have seen a similar situations before, which happens more often when the fragments are close together, however when there is some distance between them they remark that they have forgotten what they inputted in the previous situations When for each of the participants the difference in scores between two couplings and their distance in video fragments is registered, a linear regression between the two could indicate whether a correlation can be found. Performing this for every couple found no significant effects (r^2<0.01 for all couples).

Besides testing whether the distance between couplings has an effect, it could also be tested whether the order in which the couplings are displayed will have an effect. It is tested whether the videos with additional workload (through the addition of VRUs or rain) will have a different score when appearing in different order. When a participant sees the second of a pair of video fragments belonging to the same couple, they could argue that since they notice an additional variable compared to the first fragment of the couple, the video deserves a higher rating. When the same video would have been displayed before its couple, they could have given it a lower score, and give the other fragment an even lower score as a means of compensation. To test this effect, two tests are performed. It is tested whether there is a difference in the individual scores depending on whether the situation was shown before its couple, and it is tested whether the rating score difference between the two videos for each participant changes based on which situation was shown first.

For the individual videos it was tested whether an effect of the order of the couplings could be found. The individual scores for each of the extra videos in which they were displayed before their couple was compared to the scores for which they were shown second. A rANOVA shows no significant main effect of the order of the extra videos on the individual scores (p=0.681 for the VRU couples, p=0.491 for the weather effect couples and p=0.324 overall). Pairwise comparisons show no significant effects.

In order to test whether the difference between the present or not-present situations changes between the different orders of appearance, the difference in scores for each of the couples where the present situation was shown first is compared to the difference in scores for the couples where the not-present situation was shown first. A rANOVA shows no significant main effect of the order of the extra videos (p=0.347 for the VRU couples, p=0.677 for the weather effect couples and p=0.778 overall). Pairwise comparisons show only a single significant effect in one of the VRU couples (p=0.048).

6.2.3 Influence of learning effects

Even though the participants received an introduction on the use of the scale, starting the experiment and using the RSME scale may still require some getting used to. Especially for the first couple of fragments the participant is often still building a frame of reference on which they base they subsequent scores, which means it may be possible to find a . In order to find out if a learning effect can be found, it is tested whether the answers that were given at latter parts of the experiment provided different scores than the answers given at the first couple of videos. This is done by performing two paired Student's t-tests, one for which the average and standard deviation of the first 16 video fragments are compared to the average and standard deviation of the last 16 videos, and one where the average and standard deviation of the first 5 videos is compared to the average and standard deviation of the last 5 videos. The credibility of these tests is based on the assumption that over the entire sample the high and low workload situations are evenly distributed over the earlier and latter parts of the experiment. The justification for this assumption is provided in appendix F.

The results of the paired Student's t-tests showed no significant difference for either comparing the first 16 video fragments to the last 16 video fragments (p=0.767 for the mean and p=0.25 for the standard deviation) or comparing the first 5 video fragments to the last 5 video fragments (p=0.93 for the mean and p=0.33 for the standard deviation). Pearson's-r tests showed a high correlation between the means of the first five fragments and the last 27 fragments (r=0.730) as well as the means of the first 16 fragments and the last 16 fragments (r=0.896). These results do not indicate the presence of a learning effect.

6.3 Chapter Summary

In this chapter an evaluation was given on the measurement method. The method is judged on its test validity and the experiment on its experimental validity. There are two reasons which could result in a low degree of test validity. Only a self-report measure is applied, while in workload research generally two or more measurement methods are used. Self-report measures are subject to bias, as people may overestimate their driving ability. Furthermore it is unknown what the effect is of having the subject watch video fragments of traffic situations, instead of having them participate in traffic themselves. These two effects may amplify each other, since a lack of driving skill from the

participant is not necessarily correlated to a decrease in the driver's performance when using this method.

An approximation of the test validity is made by comparing the effects which are found in prior studies which employ methods of obtaining workload which contain better validity. Influence of density, HGVs and age on mental workload were found, which is in line with the results obtained through this method. No information was found on the effect of the number of lanes, which neither confirms nor denies the results found here.

For the experimental validity, there are a couple of concerns with internal and external validity. The lack of control over the variables and presence of confounding variables may cause a reduction in internal validity. Because of the nature of naturalistic driving, there is no control over the traffic situations in which the vehicle drives and the variables that occur during these situations. Variable control during the creating of the video fragments was mostly focused on the prevention of the confounding caused by road curvature and maneuvers such as overtaking. Compromises have been made mostly on the presence of roadside distractions, especially in sparingly occurring combinations of variables. A series of rANOVAs was performed to test the influence of confounding, by instead of using average scores of the two identical situations use different combinations of single video scores. The results show some difference in the statistical significance of interaction effects that were found, however the main effect of density, HGVs and age the results are still significant for each combination.

For the external validity a number of concerns are identified. Information on a number of personal characteristics were missing, such as social economic status or marital status, of which it would be preferred to have an equal distribution on the sample. Furthermore a bias is likely found towards a higher level of education, since a lot of participants were recruited among employees and students at the VU university in Amsterdam. Recruitment was centered around the cities The Hague and Amsterdam, which results in a high prevalence of participants whom are likely familiar driving in city surroundings. The possible bias introduced by this is however not testable without performing a similar study using a sample with different characteristics.

As a result of the use of a repeated measures design, some bias may be created in the results. A number of tests were performed in order to test if any order- or learning effect can be demonstrated. No significant effects were found as a result of differences in the difficulty of the first video, the order and distance between two coupled video fragments or learning effects.

7 Conclusion and recommendations

In this chapter the conclusions which are drawn from the previous two chapters are provided. Since the previous two chapters have focused on two different aspects; the results of the experiment and an evaluation of the method, as such this chapter is divided in two parts. The first part of this chapter will focus on the results of the experiment and the second part will raise concerns about these conclusions and make recommendations on how the results could be improved or validated.

7.1 Subjective workload in driving situations

The objective of the experiment was to identify predictors of subjective workload in motor way situations. A factorial analysis found a significant main effect of traffic density, presence of Heavy Goods Vehicles (HGVs) and driver's age, and no significant main effect of the number of lanes on driver's mental workload. Furthermore, a number of interaction effects between the variables were found. At lower levels of density, the presence of HGVs has an increased effect on mental workload. The younger drivers indicated requiring more effort driving at an increased number of lanes at high density, contrary to the older groups which require lesser effort. Furthermore the younger group required a relatively bigger increase in overall effort at higher density levels.

In a number of extra videos the main effect of the presence of vulnerable road users (VRUs) and adverse weather, as well as their interaction effects with age were studied. From the analysis it followed that a significant main effect for both the presence of VRUs and adverse weather was present, as well as an interaction effect between presence of VRUs and age. This showed that younger drivers found the presence of VRU in traffic resulted in more effortful driving. Post hoc analysis of the data showed no significant main effect of the driver's gender or yearly kilometrage on mental workload.

7.2 Showing video images to determine subjective workload

Chapter 5 showed some promising results, however the results themselves are only as good as the experiment method which was used to obtain them. In Chapter 6, the possible shortcomings of this research were discussed. Test validity is unknown since no workload was measured using more conventional measurement methods, disallowing a comparison. The most that could be done is see whether the results of this experiment agree with the results of studies performed in the past, which use methods with proven validity for mental workload research. A review of the relevant literature did show similar main effects being found as the ones in this experiment for main effects for density (e.g. Dingus et al., 1989; Brookhuis et al., 1991; Zeitlin, 2005; De Waard et al., 2008), age (e.g. Verwey, 2000;Makishita and Matsunaga, 2008; Cantin et al., 2009) and presence of HGVs (De Waard et al. 2008, 2009). Regardless, the results of this study should not be taken for granted until further validity is obtained through additional research. The two main aspects affecting the test validity of the measurement method are the fact that only a subjective measurement method is used, and what the bias is resulting from the fact that the participant is not the person actually driving through the situation.

A cause of a reduction of internal validity was the lack of control over the variables in the experiment. This can be improved in two ways: either assume greater control over the present

variables by setting up the conditions in which the video fragments are made, or increase the number of video fragments for each of the traffic situations, in order to reduce the effect of the confounding variables present in individual video fragments.

Other than the validity, some artefacts in the experimental method which may introduce a bias to the results have been studied. The study found no significant bias as a result from the effect of the first video, random order or learning effects. While this does not necessarily mean that the method is free of any bias, there is yet no reason to believe in any interference caused by any of these effects.

In order to attach additional validity to the results that were found, a follow-up study could be performed which relates the result of an experiment using the measurement method developed in this research to methods which are traditionally used in mental workload research. In this research one group of participants could drive the naturalistic driving car through a predetermined route using a measurement method which continuously measures their workload, for instance a PDT. The video images resulting from those trips are then converted into video fragments similar to those used in this research and would be shown to a different group of participants whom would indicate their perceived effort on a self-report scale. The results of the two tests could then be compared to each other, hopefully resulting in a significant correlation between the two.

7.3 Advantages of the method and future application

To counteract some of the negative aspects of the measurement method that have been discussed, some attention is paid to advantages of this method and its potential as a workload measurement method. The major advantage of this method when compared to other workload measurement methods is the ease of which it can be applied. Any naturalistic driving study could include this design into their research without any additional monetary cost other than possibly a small compensation for participation. Because of the short length of each participation, creating a large sample size can be done in a relatively short time period. One advantage specific to this method is the number of different situations in which the participants can be placed in a short time. The experiment in this research used a factorial repeated measures design with 3 within subjects variables, as well as an additional design with another two within subject variables and 4 stimuli per variable, all within a timespan averaging only 15 minutes per participant. While the validity of the method is yet to be determined, the experiment in this research did manage to reproduce the main effects found in prior experiments.

Additional use of the method in naturalistic driving studies could give insight on the reliability and bandwidth of the method and possibly some test validation in case correlations can be found between the results of the method and other methods employed in the research. When employing this method some attention needs to be paid on the way the experiment is set up. Especially considering the interpretation of the term effort, attention needs to be paid on giving each of the participants the same introduction. The term effort has, during this experiment, been interpreted in two different ways. A portion of the participants sees at as a scale in which different levels of effort are possible which differ in relative intensity, which is how it is intended to be interpreted during this research. The other portion of the participants sees effort as an absolute, meaning that a situation is either effortful, or not effortful at all. Here effort is seen as synonymous with high effort, and no effort with absolutely no effort. Since a lot of the situations in this experiment require little to moderate effort, these participants will group all the situations in the little or no effort category, and

reserve their higher scores for a very high effort situation, which will never appear. For this it is important that the instructions make clear that the situations are never extremely intensive and relative differences between the situations are important. It also needs to be included that in traffic it should never be possible to require absolutely no effort, which is synonymous with sleeping behind the steering wheel. It is advised to keep using a repeated measures design, as a rANOVA holds the variations between subjects into account in determining the effects. Using a random order as a means of counterbalancing should limit the influence of order effects which have not been found to cause a significant bias in the results (see paragraph 7.2).

In the end it is important to review whether the research questions mentioned in the introduction are answered. These research questions are:

1. What is mental workload and how can it be measured?
2. What aspects of car driving can result in increased mental workload?
3. How can video fragments be used to measure mental workload?
4. What is the validity of the measurement method?
5. How does the measurement method compare to other workload measurement methods?

The first three questions are answered in chapter 2, 3 and 4 respectively. The fourth question is answered in chapter 6. To answer the final research question, we focus on the measurement evaluation criteria (O'Donnel and Eggemeier, 1986; De Waard, 1996) described in paragraph 2.2. The sensitivity of the method determines how differences in workload are measurable at different levels of performance. While initially having expressed doubts about the sensitivity of the method, it is actually found to be surprisingly sensitive. This is indicated by the high number of effects which were found, even though the driving difficulty in all situations was low to moderate at best and driving performance was always (near) optimal. Diagnosticity relates to the existence of multiple aspects of workload and to what extend the method can distinguish between these. In the current research, diagnosticity is nonexistent, since a uni-dimensional scale is used. When looking at the primary task intrusion, it is important to elaborate what the primary task in this experiment is. While normally the task would be to safely operate the vehicle, since no vehicle is actually being operated by the participant it is natural to assume that there is no primary task to intrude upon. However a point could be made that the primary task of the experiment is to place yourself into the drivers position, to which there is a definite chance of intrusion. Having the rating scale appear in between two subsequent situations at very short time intervals could affect the participant's ability to empathize with each situation, considering that the participant is already thinking about what to fill in into the scale while looking at the video fragments. A way to improve this is to increase the length of the video fragments as well as implement a time period in between the RSME and the subsequent video, whereas the experiment in the current research was self-paced. The implementation requirements have already been discussed in the start of this paragraph. The low cost and time investment required to implement this method, especially when considered with conventional workload measurement methods, are the primary reason to employ this method. Operator acceptance of the method is mixed. Participants are eager to enlist themselves into the experiment, since the time requirement is very short and a compensation is offered. However the participant is not always convinced of the value of the experiment and often wonder why there are no really difficult situations included, which was most noted among the middle-aged and elderly participants. It is unknown whether this has any effect on the results however it is imaginable that this results in them

placing their scores closer to each other at the lower spectrum of the scale. The selectivity of the research determines relate to the extent to which the method actually measures mental workload. This is strictly related to the construct validity of the method, which is extensively discussed in paragraph 7.1. An actual estimate of the selectivity cannot be made, since this requires a comparison with a method of which it has been found to actually measure workload. The bandwidth and reliability of the method is the last measurement criteria and determines whether the method is applicable in different setting and performance ranges and whether repeated applications results in similar results. The importance of the instructions given to participants has been highlighted previously in this chapter and application of this method using different instructions or other modifications to the procedure could possibly have an impact on the results. Since this is the first time this method has been applied for determining workload, no information is however known on this account.

The advice made by the writer is to include this measurement method whenever possible in naturalistic study to driver's mental workload. Not only is the method easy to implement, it has shown its ability to reproduce effects which were found in studies performed in the past. Subsequent use of the method would help to give an insight into the validity and usability of the scale and possible adjustments which could improve the method. The use of this method is possible in any study which uses camera equipped vehicles. The method is especially useful in ND studies where daily occurring trips are made, which show only slight alterations (e.g. the extra videos in this research). These situations have better control over the occurring variables, which should limit confounding. In order to obtain even better variable control it is also possible to manually set up the driving scenarios, however this would greatly increase the implementation requirement, which is the main appeal of the use of this method. With the use of this method is it possible to examine any variable observable through the front view of a driver, as long as it is perceivable in the visible range of the participant. This excludes the study of in-vehicle technologies, of which a few examples were described in paragraph 3.4. Since the participant is not actively using the vehicle, effects of the use of for instance ACC will not be noticeable to the participant. It is possible to have a separate screen display the driver operating an in-vehicle technology during driving, however applying this method does not accurately convey the dual-task paradigm, as the participant will still be performing only one task (watching the screen). In addition to the study of mental workload, possible alteration could be made to the method to serve other purposes. An example is asking the participant for the contents of a traffic sign in situations with ranging environment complexity, which could be used as a driver distraction measurement method.

References

Alm, H., and Nilsson, L. (1994). Changes in driver behaviour as a function of handsfree mobile phones—a simulator study. *Accident Analysis and Prevention*,*26*(4), 441-451.

Brookhuis, K. A., van Driel, C. J., Hof, T., van Arem, B., and Hoedemaeker, M. (2009). Driving with a congestion assistant; mental workload and acceptance. *Applied ergonomics*, *40*(6), 1019-1025.

Brookhuis, K. A., de Vries, G., and de Waard, D. (1991). The effects of mobile telephoning on driving performance. *Accident Analysis and Prevention*, *23*(4), 309-316.

Cain, B. (2007). *A review of the mental workload literature*. Defense Research and Development Toronto.

Cantin, V., Lavallière, M., Simoneau, M., and Teasdale, N. (2009). Mental workload when driving in a simulator: Effects of age and driving complexity. *Accident Analysis and Prevention*, *41*(4), 763-771.

Crundall, D. E., and Underwood, G. (1998). Effects of experience and processing demands on visual information acquisition in drivers. *Ergonomics*, *41*(4), 448-458.

De Craen, S., Twisk, D. A., Hagenzieker, M. P., Elffers, H., and Brookhuis, K. A. (2007). Overestimation of skills affects drivers' adaptation to task demands. In*Proceedings of the 4th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design, Stevensen, Washington, USA* (pp. 39-45).

De Waard, D. (1996). *The measurement of drivers' mental workload*. Groningen University, Traffic Research Center.

De Waard, D., Dijksterhuis, C., and Brookhuis, K. A. (2009). Merging into heavy motorway traffic by young and elderly drivers. *Accident Analysis and Prevention*,*41*(3), 588-597.

De Waard, D., Kruizinga, A., and Brookhuis, K. A. (2008). The consequences of an increase in heavy goods vehicles for passenger car drivers' mental workload and behaviour: a simulator study. *Accident Analysis and Prevention*, *40*(2), 818-828.

Dijksterhuis, C., Brookhuis, K. A., and De Waard, D. (2011). Effects of steering demand on lane keeping behaviour, self-reports, and physiology. A simulator study. *Accident Analysis and Prevention*, *43*(3), 1074-1081.

Dingus, T. A., Hulse, M. C., Antin, J. F., and Wierwille, W. W. (1989). Attentional demand requirements of an automobile moving-map navigation system.*Transportation research part A: general*, *23*(4), 301-315.

Edquist, J., Rudin-Brown, C. M., and Lenné, M. G. (2012). The effects of on-street parking and road environment visual complexity on travel speed and reaction time. *Accident Analysis and Prevention*, *45*, 759-765.

Freund, B., Colgrove, L. A., Burke, B. L., and McLeod, R. (2005). Self-rated driving performance among elderly drivers referred for driving evaluation.*Accident Analysis and Prevention*, *37*(4), 613-618.

Fukuda, K., Stern, J. A., Brown, T. B., and Russo, M. B. (2005). Cognition, blinks, eye-movements, and pupillary movements during performance of a running memory task. *Aviation, space, and environmental medicine*,*76*(Supplement 1), C75-C85.

Godley, S. T., Triggs, T. J., and Fildes, B. N. (2002). Driving simulator validation for speed research. *Accident analysis and prevention*, *34*(5), 589-600.

Green, P. (2004). Driver distraction, telematics design, and workload managers: Safety issues and solutions. Society of Automotive Engineers.

Green, P., Lin, B., and Bagian, T. (1994). *Driver Workload as a function of road geometry: a pilot experiment* (No. GLCTTR 22-91/01).

Green, P., Lin, B. T., Schweitzer, J., Ho, H., and Stone, K. (2011). Evaluation of a method to estimate driving workload in real time: watching clips versus simulated driving. *University of Michigan Transportation Research Institute (UMTRI-2011-29). Ann Arbor, MI*.

Habenicht, S., Winner, H., Bone, S., Sasse, F., & Korzenietz, P. (2011, June). A maneuver-based lane change assistance system. In *Intelligent Vehicles Symposium (IV), 2011 IEEE* (pp. 375-380). IEEE.

Hancock, P. A., Wulf, G., Thom, D., and Fassnacht, P. (1990). Driver workload during differing driving maneuvers. *Accident Analysis and Prevention*, *22*(3), 281-290.

Harms, L. (1991). Variation in drivers' cognitive load. Effects of driving through village areas and rural junctions. *Ergonomics*, *34*(2), 151-160.

Hart, S. G., and Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Human mental workload*,*1*(3), 139-183.

Heger, R. (1998). *Driving behavior and driver mental workload as criteria of highway geometric design quality* (No. E-C003).

Hill, J. D., and Boyle, L. N. (2007). Driver stress as influenced by driving maneuvers and roadway conditions. *Transportation Research Part F: Traffic Psychology and Behaviour*, *10*(3), 177-186.

Hoedemaeker, M., and Brookhuis, K. A. (1998). Behavioural adaptation to driving with an adaptive cruise control (ACC). *Transportation Research Part F: Traffic Psychology and Behaviour*, *1*(2), 95-106.

Hogema, J.H., and Veltman, J.A. (2002). Werkbelasting en rijgedrag tijdens duisternis: eerste veldexperiment. TNO Technische Menskunde, Soesterberg.

Hoogendoorn, R. G. (2012). Empirical research and modeling of longitudinal driving behavior under adverse conditions. TRAIL Research School.

Horberry, T., Anderson, J., Regan, M. A., Triggs, T. J., and Brown, J. (2006). Driver distraction: the effects of concurrent in-vehicle tasks, road environment complexity and age on driving performance. *Accident Analysis and Prevention*,*38*(1), 185-191.

Horrey, W. J., Lesch, M. F., and Garabet, A. (2009). Dissociation between driving performance and drivers' subjective estimates of performance and workload in dual-task conditions. *Journal of safety research*, *40*(1), 7-12.

Kirk, R. E. (1996). Practical significance: A concept whose time has come.*Educational and psychological measurement*, *56*(5), 746-759.

Konstantopoulos, P., Chapman, P., and Crundall, D. (2010). Driver's visual attention as a function of driving experience and visibility. Using a driving simulator to explore drivers' eye movements in day, night and rain driving.*Accident Analysis and Prevention*, *42*(3), 827-834.

Lin, B. T., Green, P., Kang, T. P., and Lo, E. W. (2012). Development and evaluation of new anchors for ratings of driving workload.

Ma, R., and Kaber, D. B. (2005). Situation awareness and workload in driving while using adaptive cruise control and a cell phone. *International Journal of Industrial Ergonomics*, *35*(10), 939-953.

Makishita, H., and Matsunaga, K. (2008). Differences of drivers' reaction times according to age and mental workload. *Accident Analysis and Prevention*, *40*(2), 567-575.

Martens, M. H., and Van Winsum, W. (2000). Measuring distraction: the peripheral detection task. *TNO Human Factors, Soesterberg, Netherlands*.

Michon, J. A. (1985). A critical view of driver behavior models: what do we know, what should we do? (pp. 485-524). Springer US.

Mourant, R. R., and Rockwell, T. H. (1972). Strategies of visual search by novice and experienced drivers. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *14*(4), 325-335.

Muckler, F. A., and Seven, S. A. (1992). Selecting performance measures:" Objective" versus" subjective" measurement. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *34*(4), 441-455.

Nilsson, L. (1995, November). Safety effects of adaptive cruise controls in critical traffic situations. In *Steps Forward. Intelligent Transport Systems World Congress* (No. Volume 3).

O'Donnell, R. D., and Eggemeier, F. T. (1986). Workload assessment methodology.Wickens 1984

Parkes, A. M., Ashby, M. C., and Fairclough, S. H. (1991, October). The effects of different in-vehicle route information displays on driver behaviour. In *Vehicle Navigation and Information Systems Conference, 1991* (Vol. 2, pp. 61-70). IEEE.

Patten, C. J., Kircher, A., Östlund, J., and Nilsson, L. (2004). Using mobile telephones: cognitive workload and attention resource allocation. *Accident analysis and prevention*, *36*(3), 341-350.

Patten, C. J., Kircher, A., Östlund, J., Nilsson, L., and Svenson, O. (2006). Driver experience and cognitive workload in different traffic environments. *Accident Analysis and Prevention*, *38*(5), 887-894.

Paxion, J., Freydier, C., Galy, E., and Berthelon, C. (2013, January). Subjective workload and performance of young drivers faced to unexpected pedestrian crossings. In *11th International Conference on Naturalistic Decision Making 2013*.

Reid, G. B., and Nygren, T. E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. *Advances in Psychology*, *52*, 185-218.

Rosey, F., Auberlet, J. M., Moisan, O., and Dupré, G. (2009). Impact of Narrower Lane Width. *Transportation Research Record: Journal of the Transportation Research Board*, *2138*(1), 112-119.

Schaap, T. W., Horst, A. R. A., Arem, B. V., & Brookhuis, K. A. (2009). The relationship between driver distraction and mental workload.

Schmider, E., Ziegler, M., Danay, E., Beyer, L., & Bühner, M. (2010). Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *6*(4), 147.

Schweitzer, J., and Green, P. E. (2007). Task acceptability and workload of driving city streets, rural roads, and expressways: Ratings from video clips.

Stern, J. A., Boyer, D., and Schroeder, D. (1994). Blink rate: a possible measure of fatigue. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *36*(2), 285-297.

Teh, E., Jamson, S., Carsten, O., and Jamson, H. (2014). Temporal fluctuations in driving demand: The effect of traffic complexity on subjective measures of workload and driving performance. *Transportation Research Part F: Traffic Psychology and Behaviour*, *22*, 207-217.

Törnros, J., and Bolling, A. (2006). Mobile phone use–effects of conversation on mental workload and driving speed in rural and urban environments.*Transportation Research Part F: Traffic Psychology and Behaviour*, *9*(4), 298-306.

Verwey, W. B. (1989, September). Simple in-car route guidance information from another perspective: modality versus coding. In *Vehicle Navigation and Information Systems Conference, 1989. Conference Record* (pp. 56-60). IEEE.

Verwey, W. B. (2000). On-line driver workload estimation. Effects of road situation and age on secondary task measures. *Ergonomics*, *43*(2), 187-209.

Vollrath, M., Schleicher, S., and Gelau, C. (2011). The influence of Cruise Control and Adaptive Cruise Control on driving behaviour–A driving simulator study. *Accident Analysis and Prevention*, *43*(3), 1134-1139.Kick 1996

Ward, N. J., Fairclough, S., and Humphreys, M. (1995). The effect of task automatisation in the automotive context: A field study of an Autonomous Intelligent Cruise Control system. *Proceedings of the International Confeence on Experimental analysis and Measurement of Situation Awareness. Daytona Beach, FL. November*, 1-3.

Wickens, C. D. (1991). Processing resources and attention. *Multiple-task performance*, 3-34.

Wierwille, W. W., and Casali, J. G. (1983, October). A validated rating scale for global mental workload measurement applications. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 27, No. 2, pp. 129-133). Sage Publications.

Wong, J. T., & Huang, S. H. (2009). Modeling Driver Mental Workload for Accident Causation and Prevention. *Journal of the Eastern Asia Society for Transportation Studies*, *8*, 1918-1933.

Wong, J. T., Chung, Y. S., & Huang, S. H. (2010). Determinants behind young motorcyclists' risky riding behavior. *Accident Analysis & Prevention*, *42*(1), 275-281.

Wu, C., and Liu, Y. (2007). Queuing network modeling of driver workload and performance. *Intelligent Transportation Systems, IEEE Transactions on*, *8*(3), 528-537.

Yerkes, R. M., and Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of comparative neurology and psychology*,*18*(5), 459-482. Hebb, D. O. (1955). Drives and the CNS (conceptual nervous system). Psychological review, 62(4), 243.

Young, M. S., and Stanton, N. A. (2004). Taking the load off: investigations of how adaptive cruise control affects mental workload. *Ergonomics*, *47*(9), 1014-1035.

Zeitlin, L. R. (1995). Estimates of driver mental workload: A long-term field trial of two subsidiary tasks. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *37*(3), 611-621.

Zijlstra, F. R. H. (1993). Efficiency in work behaviour: A design approach for modern tools(Ph. D. Thesis).

Appendix A - Video Fragments

In an attempt to better familiarize the reader with the contents of the video fragments, this appendix shows screen captures from the different video fragments. This is done in pairs, showing the identical combinations together and the opposing extra videos. The abbreviations shown in the captions describe the levels of the variables present in the video fragment. A *2* or *3* for the number of lanes, *H* for high density and *L* for low density, a *Y* for the presence of HGVs and *N* for the lack of their presence.

Appendix A.1 Motor Way fragments

Figure A.1 2HN

Figure A.2 2HY

Figure A.3 2LN

Figure A.4 2LY

Figure A.5 3HN

Figure A.6 3HY

Figure A.7 3LN

Figure A.8 3LY

Appendix A.2 Extra Videos

Again the videos are shown in pairs, with the fragment where the additional condition is not present shown on the left. In the weather effect videos the rain is not always clearly visible, therefore it is made sure that the windscreen wipers are clearly seen. The abbreviations *V* and *W* stand for presence of VRUs or weather effects, with the *Y* indicating the presence and *N* indicating no presence.

Figure A.9 V1

Figure A.10 V2

Figure A.11 V3

Figure A.12 V4

Figure A.13 W1

Figure A.14 W2

Figure A.15 W3

Figure A.16 W4

Appendix B - Tests of Normality

In this appendix, the results of the tests of normality are displayed, both with the original values and log^{10} transformations. The values in the table are p values levels, where a value of 0.05 or lower indicates significant departure from the normal distribution. All tests are performed using the Age category as a factor.

Table B.1 Normality Test Motor way Situations

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Age_cat	Statistic	df	Sig.	Statistic	df	Sig.
h2n	1,00	,145	20	,200*	,944	20	,286
	2,00	,116	20	,200*	,967	20	,701
	3,00	,099	20	,200*	,965	20	,658
h2y	1,00	,140	20	,200*	,965	20	,637
	2,00	,099	20	,200*	,955	20	,455
	3,00	,180	20	,087	,928	20	,141
l2n	1,00	,173	20	,118	,926	20	,127
	2,00	,116	20	,200*	,938	20	,224
	3,00	,148	20	,200*	,932	20	,168
l2y	1,00	,149	20	,200*	,896	20	,034
	2,00	,133	20	,200*	,951	20	,388
	3,00	,139	20	,200*	,945	20	,293
h3n	1,00	,134	20	,200*	,951	20	,381
	2,00	,147	20	,200*	,970	20	,746
	3,00	,147	20	,200*	,940	20	,237
h3y	1,00	,102	20	,200*	,955	20	,446
	2,00	,111	20	,200*	,929	20	,145
	3,00	,156	20	,200*	,931	20	,159
l3n	1,00	,262	20	,001	,849	20	,005
	2,00	,230	20	,007	,904	20	,048
	3,00	,123	20	,200*	,950	20	,370
l3y	1,00	,145	20	,200*	,924	20	,117
	2,00	,158	20	,200*	,916	20	,084
	3,00	,214	20	,017	,887	20	,023

Table B.1 shows the tests of normality for the motor way situations. As can been seen from the table not all combinations are insignificantly different from the normal distribution. A normality test on a log10 transformation on the data is performed to check whether they result in uniform normality.

Table B.7.1 Normality tests log10 transformation Motor way Situations

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Age_cat	Statistic	df	Sig.	Statistic	df	Sig.
h2n	1,00	,106	20	,200*	,937	20	,207
	2,00	,114	20	,200*	,959	20	,528
	3,00	,169	20	,137	,925	20	,126
h2y	1,00	,172	20	,121	,941	20	,254
	2,00	,112	20	,200*	,951	20	,376
	3,00	,132	20	,200*	,958	20	,504
l2n	1,00	,116	20	,200*	,979	20	,913
	2,00	,244	20	,003	,828	20	,002
	3,00	,210	20	,022	,868	20	,011
l2y	1,00	,073	20	,200*	,989	20	,997
	2,00	,141	20	,200*	,906	20	,054
	3,00	,087	20	,200*	,965	20	,649
h3n	1,00	,144	20	,200*	,909	20	,062
	2,00	,128	20	,200*	,949	20	,352
	3,00	,123	20	,200*	,952	20	,404
h3y	1,00	,176	20	,106	,890	20	,026
	2,00	,140	20	,200*	,953	20	,409
	3,00	,083	20	,200*	,981	20	,952
l3n	1,00	,200	20	,035	,939	20	,233
	2,00	,236	20	,005	,822	20	,002
	3,00	,149	20	,200*	,929	20	,146
l3y	1,00	,112	20	,200*	,981	20	,949
	2,00	,140	20	,200*	,952	20	,405
	3,00	,107	20	,200*	,977	20	,892

The resulting tests are not a big improvement on the test done without the log10 transformation, and therefore it is chosen to use the original data.

For the extra videos the same steps are repeated. Tables B.3 and B.4 display the results for the actual data and log10 transformations respectively. The abbreviations stand for VRU (v) or Weather (w), Yes (y) or No (n) for their presence and a number to indicate the couples.

Table B.7.2 Normality tests Extra videos

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Age_cat	Statistic	df	Sig.	Statistic	df	Sig.
v1n	1,00	,184	20	,074	,904	20	,049
	2,00	,094	20	,200*	,960	20	,537
	3,00	,131	20	,200*	,946	20	,316
v1y	1,00	,194	20	,048	,908	20	,059
	2,00	,167	20	,146	,953	20	,422
	3,00	,156	20	,200*	,964	20	,626
v2n	1,00	,189	20	,059	,908	20	,058
	2,00	,157	20	,200*	,948	20	,344
	3,00	,198	20	,038	,788	20	,001
v2y	1,00	,221	20	,011	,929	20	,147
	2,00	,102	20	,200*	,951	20	,378
	3,00	,169	20	,137	,891	20	,028
v3n	1,00	,139	20	,200*	,929	20	,148
	2,00	,129	20	,200*	,965	20	,642
	3,00	,237	20	,004	,848	20	,005
v3y	1,00	,146	20	,200*	,978	20	,910
	2,00	,208	20	,024	,840	20	,004
	3,00	,135	20	,200*	,960	20	,548
v4n	1,00	,092	20	,200*	,964	20	,634
	2,00	,083	20	,200*	,991	20	,999
	3,00	,134	20	,200*	,949	20	,350
v4y	1,00	,141	20	,200*	,963	20	,595
	2,00	,132	20	,200*	,942	20	,265
	3,00	,210	20	,021	,934	20	,185
w1n	1,00	,179	20	,094	,897	20	,036
	2,00	,141	20	,200*	,956	20	,459
	3,00	,112	20	,200*	,970	20	,764
w1y	1,00	,110	20	,200*	,972	20	,802
	2,00	,108	20	,200*	,953	20	,408
	3,00	,145	20	,200*	,949	20	,357
w2n	1,00	,217	20	,014	,906	20	,054
	2,00	,147	20	,200*	,944	20	,291
	3,00	,143	20	,200*	,965	20	,641
w2y	1,00	,072	20	,200*	,978	20	,900
	2,00	,155	20	,200*	,964	20	,625
	3,00	,167	20	,144	,943	20	,279
w3n	1,00	,125	20	,200*	,955	20	,455
	2,00	,167	20	,147	,930	20	,152
	3,00	,198	20	,038	,943	20	,273
w3y	1,00	,145	20	,200*	,967	20	,700
	2,00	,143	20	,200*	,937	20	,208
	3,00	,133	20	,200*	,954	20	,427
w4n	1,00	,206	20	,026	,901	20	,043
	2,00	,202	20	,032	,888	20	,024
	3,00	,212	20	,019	,888	20	,025
w4y	1,00	,173	20	,119	,916	20	,081
	2,00	,129	20	,200*	,960	20	,537
	3,00	,156	20	,200*	,937	20	,207

Table B.7.3 Normality tests log10 transformation Extra videos

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Age_cat	Statistic	df	Sig.	Statistic	df	Sig.
v1n	1,00	,134	20	,200*	,944	20	,289
	2,00	,163	20	,170	,919	20	,096
	3,00	,136	20	,200*	,944	20	,285
v1y	1,00	,169	20	,135	,905	20	,052
	2,00	,137	20	,200*	,973	20	,817
	3,00	,254	20	,002	,742	20	,000
v2n	1,00	,130	20	,200*	,958	20	,509
	2,00	,225	20	,009	,735	20	,000
	3,00	,138	20	,200*	,931	20	,162
v2y	1,00	,154	20	,200*	,927	20	,136
	2,00	,103	20	,200*	,947	20	,321
	3,00	,114	20	,200*	,940	20	,244
v3n	1,00	,095	20	,200*	,951	20	,387
	2,00	,144	20	,200*	,899	20	,039
	3,00	,250	20	,002	,830	20	,003
v3y	1,00	,164	20	,166	,855	20	,007
	2,00	,179	20	,091	,898	20	,037
	3,00	,186	20	,067	,785	20	,001
v4n	1,00	,140	20	,200*	,946	20	,315
	2,00	,217	20	,014	,721	20	,000
	3,00	,249	20	,002	,799	20	,001
v4y	1,00	,181	20	,084	,931	20	,164
	2,00	,108	20	,200*	,969	20	,733
	3,00	,139	20	,200*	,924	20	,120
w1n	1,00	,139	20	,200*	,943	20	,276
	2,00	,301	20	,000	,644	20	,000
	3,00	,265	20	,001	,798	20	,001
w1y	1,00	,129	20	,200*	,904	20	,049
	2,00	,151	20	,200*	,878	20	,016
	3,00	,238	20	,004	,870	20	,012
w2n	1,00	,143	20	,200*	,945	20	,295
	2,00	,255	20	,001	,790	20	,001
	3,00	,248	20	,002	,772	20	,000
w2y	1,00	,128	20	,200*	,959	20	,524
	2,00	,157	20	,200*	,860	20	,008
	3,00	,205	20	,027	,745	20	,000
w3n	1,00	,176	20	,104	,912	20	,068
	2,00	,172	20	,122	,863	20	,009
	3,00	,317	20	,000	,726	20	,000
w3y	1,00	,204	20	,028	,910	20	,062
	2,00	,260	20	,001	,714	20	,000
	3,00	,172	20	,121	,823	20	,002
w4n	1,00	,166	20	,151	,952	20	,401
	2,00	,196	20	,043	,897	20	,036
	3,00	,185	20	,071	,875	20	,014
w4y	1,00	,113	20	,200*	,948	20	,343
	2,00	,159	20	,199	,957	20	,485
	3,00	,135	20	,200*	,949	20	,352

The results found here are similar to the motor way situations. Therefore again the raw data is used over the log10 transformations.

Appendix C - Differences between same situations

In this appendix the differences between the situations which are combinations of the same variables are shown. This is done by performing a paired-Student's t-test between the situations as well as a Pearson's correlation tests. Results from both tests are displayed in table C.1. The abbreviations: H, L, 2, 3, N and Y stand for the density (High or Low), number of lanes (2 or 3) and presence of HGV's (Yes or No).

Pair	Sig. (2-tailed)	Correlation (r)
2-H-N	0.121	0.542
2-H-Y	0.001	0.782
2-L-N	0.001	0.677
2-L-Y	0.326	0.696
3-H-N	0.002	0.812
3-H-Y	0.154	0.758
3-L-N	0.002	0.670
3-L-Y	0.238	0.638

As can be seen from the table, there are a number of significant differences between the video pairs, which would preferably not be significantly different. The results do however show a decent to high correlation in most of the cases.

Appendix D - Single fragments tests

In this appendix the rANOVA is once again performed, however this time instead of using the averages which are calculated from the identical situations, different combinations of single fragments are used. The results are reported in the same way as in chapter 5, the values in the tables show the effect sizes obtained through the rANOVA, with a single asterisk indicating the effect is significant with a p value of 0.05 and a double asterisk indication the effect is significant with a p value of 0.01.

Table D. 1 Results using averages of fragments (values are standardized effect sizes (η^2))

	Main Effect	Density	Lanes	HGVs	Age
Density	0.749**	x	0.001	0.162**	0.121*
Lanes	0.014	0.001	x	0.035	0.142*
HGVs	0.465**	0.162**	0.035	x	0.065
Age	0.163*	0.111*	0.142*	0.065	x

Table D.1 shows the original rANOVA performed using the averages of each identical situation fragment. The subsequent tables shown in this appendix should be compared to the results shown in this table to notice any differences.

Table D. 2 Results using lowest scoring fragments (values are standardized effect sizes (η^2))

	Main Effect	Density	Lanes	HGVs	Age
Density	0.688**	x	0.002	0.220**	0.073
Lanes	0	0.002	x	0.043	0.029
HGVs	0.450**	0.220**	0.043	x	0.047
Age	0.152*	0.073	0.029	0.047	x

Table D.1 shows the result of the rANOVA where the lowest scoring version of each combination of variables is used. It can be seen that the interaction effects found with age are no longer significant in this version of the rANOVA.

Table D. 3 Results using highest scoring fragments (values are standardized effect sizes (η2))

	Main Effect	Density	Lanes	HGVs	Age
Density	0.707**	x	0.002	0.047	0.125*
Lanes	0.013	0.002	x	0.005	0.147*
HGVs	0.300**	0.047	0.005	x	0.045
Age	0.163**	0.125*	0.147*	0.045	x

Table D.2 shows the results of the rANOVA where the highest scoring version of each combination of variables is used. In this version the interaction effects with age are again significant, however the density x HGV interaction effect, which was highly significant in the other version is no longer significant.

Table D. 4 Results using smallest differences between situations (values are standardized effect sizes (η2))

	Main Effect	Density	Lanes	HGVs	Age
Density	0.725**	x	0.003	0.369**	0.107*
Lanes	0.004	0.003	x	0.035	0.153*
HGVs	0.298**	0.369**	0.035	x	0.061
Age	0.175**	0.107*	0.153*	0.061	x

Table D.3 shows the results of the rANOVA where the differences between the situations in close proximity in difficulty are put as close together as possible. This is done by choosing first the situations for which the subtracted means are as small as possible and then adding the remaining ones by proximity to the chosen ones. In this example the number of significant effects found is the same as in the original situations

Table D. 5 Result using greatest differences between situations (values are standardized effect sizes (η2))

	Main Effect	Density	Lanes	HGVs	Age
Density	0.732**	x	0.008	0.103*	0.084
Lanes	0.021	0.008	x	0.035	0.085
HGVs	0.315**	0.103*	0.035	x	0.053
Age	0.139*	0.084	0.085	0.053	x

Table D.4 shows the results of the rANOVA where the differences between the situations who are in close in proximity are chosen to be as large as possible. Here the age interaction effects are once again not found.

In the end it likely does not matter which combinations are used, it is expected to find some differences in statistical significances with mostly all possible combinations. One thing that does remain stable however is the occurrence of the main effects of density, presence of HGVs and age, which are significant in every combination.

The conclusion which can be made from this test is that the differences in the individual fragments are sizeable enough to make a difference in the results when taken separately. In order to provide better results in the future, two recommendations are made on this account. The first option is to obtain better control over the confounding variables in the individual fragments, by setting up the situations instead of selecting fragments from an existing database. When this is done the differences between the videos should be smaller and thus smaller differences should be found when comparing the results. The second option is to include more than two fragments for each situation, so that an average over a larger number of videos can be taken. This should reduce the impact of the individual differences between the identical situations.

Appendix E - Scores

In this appendix the results of the experiment are shown, showing the age, gender, driving experience, number of times the participants uses the motor way, yearly kilometrage and the scores which are given to each video fragment. The abbreviations: H, L, 2, 3, N and Y stand for the density (High or Low), number of lanes (2 or 3) and presence of HGV's (Yes or No). For the extra videos, the abbreviations *V* and *W* stand for presence of VRUs or weather effects, with the *Y* indicating the presence and *N* indicating no presence.

Table E.1 Scores motor way situations

Age	Gender	Experienc	Motorway	Km/year	2hn1	2hn2	2hy2	2hy2	2ln1	2ln2	2ly1	2ly2	3hn1	3hn2	3hy1	3hy2	3ln1	3ln2	3ly1	3ly2
18	m	1	3	1000	74	58	63	90	26	14	25	28	71	43	63	71	15	16	36	19
20	v	2	3		77	61	43	71	20	30	34	34	56	61	76	83	49	35	57	25
20	v	2	4	400	44	74	38	43	35	26	72	56	71	57	47	56	37	26	71	58
21	v	3	6		29	45	64	67	13	12	26	16	57	41	69	68	14	12	39	16
21	m	2	3	1000	26	26	37	36	10	14	25	25	27	14	25	38	12	12	26	26
22	m	2	4	800	45	55	40	53	17	24	28	37	44	43	57	40	23	20	34	34
22	v	4	5		57	71	45	38	26	19	15	19	71	38	49	75	14	14	14	16
22	m	4	2	1200	25	26	32	46	14	13	26	26	45	26	37	57	15	18	23	26
23	v	4	5	3000	26	23	30	35	10	12	15	18	36	29	22	17	10	13	19	18
23	m	4	6	15000	51	35	21	21	15	17	17	26	20	26	56	49	18	12	19	17
23	m	5	4	4000	53	55	35	44	27	31	26	34	60	70	40	38	25	26	44	44
24	v	6	10	1000	20	32	42	42	14	37	39	33	58	55	58	49	33	28	37	22
24	m	5	8		30	31	22	24	12	35	34	24	35	33	35	44	15	18	33	43
24	v	3	2	2000	53	22	57	36	12	19	15	24	56	40	101	49	19	12	15	40
24	v	6	15	6500	65	75	74	68	28	60	58	42	70	67	84	74	39	27	52	66
25	v	5	30		64	31	61	70	37	33	51	33	71	55	84	58	23	15	35	33
25	m	5	6	2000	14	25	14	34	13	12	11	13	35	27	25	13	13	12	13	35
25	m	7		5200	32	26	38	40	12	18	25	18	41	20	56	42	12	14	18	18
25	v	7	40	10000	67	84	37	71	14	12	18	30	54	20	47	26	15	14	14	56
25	m	6	6		7	15	14	13	7	7	14	13	13	13	14	15	7	9	8	13
31	m	12	10		37	29	34	39	14	21	22	12	48	42	35	43	16	12	16	28
32	m	13	20		22	60	10	16	9	10	11	11	33	12	12	16	10	9	12	13
33	m	13	2	10000	16	32	26	25	13	14	22	19	24	20	24	28	14	12	25	14
38	v	19	6		33	37	29	38	18	28	26	29	37	36	37	29	28	25	26	36
38	m	20	5	5000	16	11	19	22	1	5	7	8	23	7	12	19	8	7	2	13
41	v	21	3	7000	28	12	11	23	4	6	9	10	12	11	12	12	9	5	11	6
41	m	21	8	8000	15	34	22	17	2	5	17	13	25	34	11	11	5	3	11	7
42	v	23	2		25	30	35	47	13	18	17	21	38	37	36	38	13	14	21	18
43	v	15	2	5000	38	46	46	43	25	25	46	41	48	56	56	36	26	21	34	41
44	m	24	8	25000	12	70	16	37	8	6	12	12	12	17	25	19	7	8	12	18
45	m	25	7	7000	20	25	19	30	13	15	28	18	27	23	37	29	12	12	21	24
45	v	23	2	5000	47	20	37	18	15	12	37	34	37	32	51	41	13	13	13	13
47	v	25	2		55	62	53	43	13	21	37	54	59	42	68	64	11	8	46	31
47	m	20	8	6000	34	45	30	42	15	21	27	36	37	25	35	34	16	12	32	27
47	v	29	4	5000	14	24	15	13	1	1	3	1	25	16	29	19	1	1	1	12
48	m	29	7	35000	14	13	25	38	1	1	12	13	13	37	36	18	1	1	13	12
49	m	30	20	40000	70	37	36	56	13	12	26	12	25	25	26	26	12	13	12	13
49	v	23	20	15000	13	25	26	25	12	27	14	24	25	26	25	25	12	12	13	26
50	m	25	8	12000	1	14	25	4	1	1	1	5	3	12	6	18	12	2	13	1
50	m	30	4	20000	29	24	25	38	22	30	34	18	46	36	31	26	31	24	21	28
66	m	45	10		26	25	49	86	6	1	43	21	44	45	43	61	1	26	43	38
66	m	48	40	15000	32	32	33	26	15	22	24	21	23	31	33	36	20	17	22	22
66	v	47	10	10000	13	12	11	37	37	37	37	11	12	1	3	13	12	2	2	12
67	m	49	20	10000	25	26	25	25	12	25	26	12	26	25	25	13	36	12	25	25
67	m	47	6	16000	12	71	47	71	1	1	40	37	50	37	26	24	13	1	25	14
67	m	49	20	10000	57	26	36	38	11	13	25	26	24	37	24	55	12	12	25	25
67	v	48	20	2000	37	56	38	38	26	26	25	25	13	13	39	25	26	13	26	25
67	v	45	20	8000	11	12	12	12	3	12	11	13	12	11	24	12	4	2	2	18
68	m	45	8	16000	37	36	35	24	24	13	23	12	24	35	23	23	22	11	23	23
69	v	49	2		38	33	27	23	6	19	14	10	20	15	13	33	12	8	18	15
71	m	40	4	15000	49	25	57	52	13	12	22	16	20	18	37	38	11	13	19	23
73	m	50	4	5000	38	38	40	57	13	26	42	27	51	55	38	68	28	14	38	27
74	v	55	2	6500	56	56	70	69	24	70	38	56	56	69	83	70	38	26	56	70
74	m	55	10	10000	65	65	72	102	38	40	46	63	74	75	87	84	51	14	39	81
75	m	54	5	14000	41	56	47	63	24	44	37	26	50	38	57	55	25	26	39	39
77	m	59	10	8000	42	64	23	46	8	38	12	12	31	25	52	39	38	12	15	22
77	m	48	5	15000	24	28	57	67	48	36	24	68	46	33	67	22	36	54	43	27
78	v	44	2	12000	19	16	27	24	10	13	15	14	20	24	25	24	12	11	13	16
81	m	54	10	25000	6	18	14	30	4	4	11	10	13	5	18	13	5	2	7	10
87	v	66	4	10000	1	37	24	35	1	1	9	1	11	2	24	1	2	1	2	27

Table E.2 Scores extra videos

Age	Gender	Experienc	Motorway	Km/year	v1n	v1y	v2n	v2y	v3n	v3y	v4n	v4y	w1n	w1y	w2n	w2y	w3n	w3y	w4n	w4y
18	m	1	3	1000	29	41	25	90	39	66	52	74	26	58	57	39	69	45	10	18
20	v	2	3		57	63	65	65	33	79	70	58	29	77	37	60	49	37	23	63
20	v	2	4	400	71	78	57	39	64	57	47	71	46	80	71	46	63	50	56	70
21	v	3	6		20	42	28	38	17	46	32	51	36	55	33	45	20	22	25	36
21	m	2	3	1000	36	37	14	58	26	71	37	71	13	38	26	56	37	14	14	13
22	m	2	4	800	32	38	54	81	52	72	57	66	77	60	50	83	42	46	36	40
22	v	4	5		25	32	26	38	22	46	14	38	27	38	25	47	41	60	14	44
22	m	4	2	1200	29	70	48	41	46	58	47	80	40	46	26	38	53	35	70	35
23	v	4	5	3000	16	18	22	33	19	51	25	57	31	45	20	36	33	37	10	24
23	m	4	6	15000	63	73	28	38	31	52	56	69	31	82	37	49	52	44	13	24
23	m	5	4	4000	58	74	58	61	31	57	72	74	23	69	60	54	61	62	39	38
24	v	6	10	1000	37	43	50	58	49	59	54	66	57	53	53	57	64	50	38	57
24	m	5	8		67	73	29	38	30	43	22	69	39	92	66	75	48	47	36	23
24	v	3	2	2000	23	43	15	30	13	34	88	112	12	14	29	51	51	44	13	16
24	v	6	15	6500	53	71	85	85	62	94	63	84	70	80	70	79	54	68	43	69
25	v	5	30		50	84	32	76	44	85	84	113	39	93	35	63	57	62	25	76
25	m	5	6	2000	22	37	37	12	13	14	22	51	12	22	26	28	22	20	12	24
25	m	7		5200	36	46	20	32	18	48	41	41	28	47	32	26	38	44	15	26
25	v	7	40	10000	26	48	16	16	14	58	82	89	29	64	85	71	26	34	47	49
25	m	6	6		20	38	13	30	28	40	32	24	24	34	19	21	24	26	6	13
31	m	12	10		35	23	39	42	23	29	38	48	19	32	31	43	54	37	16	27
32	m	13	20		30	34	23	26	14	25	23	32	28	61	24	26	12	15	9	9
33	m	13	2	10000	21	39	23	49	43	37	40	40	34	42	22	37	19	37	12	15
38	v	19	6		31	36	27	38	34	53	26	39	25	56	26	54	38	52	27	33
38	m	20	5	5000	18	32	13	29	12	37	32	24	26	37	10	31	11	10	7	24
41	v	21	3	7000	11	13	21	51	13	25	22	26	22	13	34	32	14	22	11	17
41	m	21	8	8000	12	25	22	28	8	23	11	31	13	21	20	22	14	21	4	10
42	v	23	2		23	24	16	20	20	25	21	38	15	23	25	38	27	22	14	26
43	v	15	2	5000	54	66	40	36	46	52	55	56	54	67	51	60	53	53	40	41
44	m	24	8	25000	44	56	37	56	28	27	18	23	25	25	57	18	26	25	18	16
45	m	25	7	7000	27	50	38	31	25	30	36	57	40	44	37	38	60	35	12	30
45	v	23	2	5000	31	45	18	18	31	26	33	52	25	69	26	38	26	35	12	56
47	v	25	2		59	59	42	44	25	31	48	82	23	70	60	72	51	56	40	44
47	m	20	8	6000	38	42	37	48	36	54	35	53	39	54	52	45	36	32	16	37
47	v	29	4	5000	14	19	3	20	10	25	14	16	1	49	1	30	14	23	1	14
48	m	29	7	35000	32	32	19	14	20	25	28	31	30	48	23	26	26	28	12	29
49	m	30	20	40000	36	25	25	56	56	37	25	56	37	70	57	71	56	56	25	36
49	v	23	20	15000	13	25	12	25	13	26	12	13	24	25	12	25	25	25	12	24
50	m	25	8	12000	4	20	1	13	2	19	1	17	1	6	3	5	2	1	5	12
50	m	30	4	20000	55	50	28	32	43	45	33	25	28	33	36	53	34	56	28	33
66	m	45	10		72	69	10	38	32	27	63	79	37	73	38	70	37	65	51	35
66	m	48	40	15000	37	31	28	35	32	42	35	48	35	63	37	38	37	38	27	38
66	v	47	10	10000	11	2	12	12	2	11	2	12	2	11	2	11	2	2	1	47
67	m	49	20	10000	26	25	24	26	26	26	12	37	25	25	25	37	37	36	25	25
67	m	47	6	16000	48	57	12	69	27	72	40	69	33	66	37	56	69	76	10	13
67	m	49	20	10000	24	26	25	24	25	25	35	37	26	36	37	25	25	38	12	25
67	v	48	20	2000	39	37	36	55	36	23	37	57	24	55	25	55	37	26	14	38
67	v	45	20	8000	12	12	13	11	11	12	4	25	3	26	12	25	24	24	1	12
68	m	45	8	16000	35	35	22	24	23	37	24	35	24	55	23	23	23	24	34	36
69	v	49	2		33	35	12	19	29	37	39	28	20	39	22	32	34	32	5	30
71	m	40	4	15000	61	69	38	56	26	57	47	57	48	80	44	64	38	49	13	57
73	m	50	4	5000	42	57	34	40	44	41	27	55	46	57	50	49	55	56	14	41
74	v	55	2	6500	84	83	71	69	83	56	69	84	56	83	56	70	55	69	56	84
74	m	55	10	10000	58	67	80	69	76	55	71	79	68	100	60	71	75	72	50	74
75	m	54	5	14000	49	56	38	16	25	64	35	25	41	70	39	55	56	56	38	56
77	m	59	10	8000	24	37	14	33	37	41	41	37	25	71	26	57	37	55	10	70
77	m	48	5	15000	38	53	24	49	35	55	27	38	36	84	53	61	25	25	48	41
78	v	44	2	12000	28	35	25	19	13	13	22	33	16	24	17	24	35	16	20	23
81	m	54	10	25000	28	18	17	18	35	13	14	34	10	7	8	39	6	16	4	17
87	v	66	4	10000	12	1	11	13	12	1	1	11	1	13	1	2	1	14	1	12

Appendix F - Assumption test

In order to validate the results from the learning effects test, it is checked whether the assumption that the videos are evenly distributed over the first 5 videos as a result of the random ordering is correct. At first this is done by looking at the distribution of the videos of over the first five videos, when the videos are uniformly distributed over the participants the assumption holds true. Figure F.1 shows the distribution.

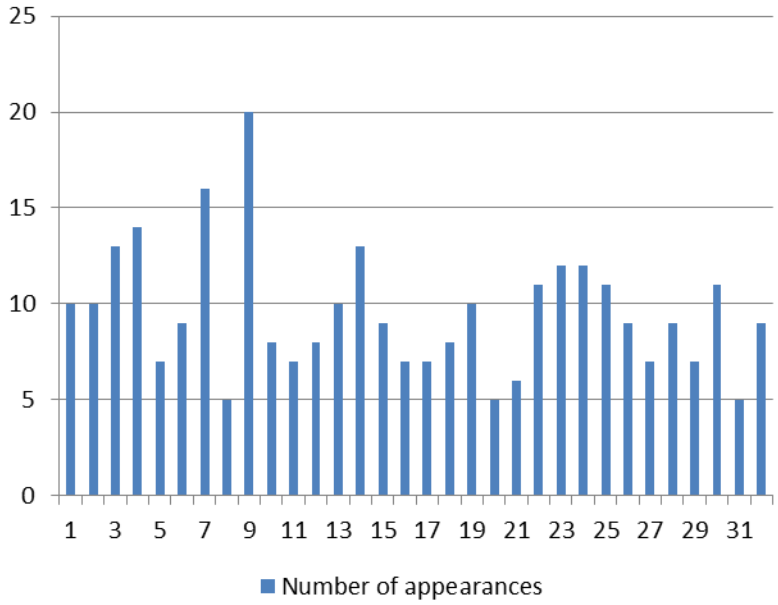


Figure F.1 Number of appearances of the individual videos over the first 5 videos

In the graph the numbers on the horizontal axis represent the individual video fragments. And the numbers on the vertical axes the number of times they appear in the first 5 videos over all the participants. From a first glance, the distribution is not necessarily uniformly distributed enough to be able to be able to perform the test. It is however not necessary for the videos to be equally represented, as long as there is not a large difference in difficulty present. When two traffic situations are considered to be equally difficult it should make no difference in what quantities they are present in the first five fragments. Therefore it may be sufficient to see if overall the first and last five situations are equally difficult as the rest of the situations. As a measure of difficulty, the situations are again assigned the average score they received over all participants. A paired t-test performed between the scores of the first 5 video fragments and the scores of the other 27 fragments shows no significant difference ($p=0.47$). As a result of this it is decided that the assumption that the first 5 videos fragments over all the participants are equal in difficulty as the rest of the video fragments justified.

Appendix G - Comparison with earlier performed study

At the start of this research it was assumed that the method which is used in this research, letting subjects judge video images on perceived effort rather than having them drive themselves, was never used before in workload measurement research. Therefore a lot of attention has been spend towards developing the method in order to have it best serve its purpose. Towards the end of the research period however, it was found that a method very similar to the current research had already been developed. Furthermore some validation studies had already been performed. In this appendix some comparisons are drawn between the two methods and the validity obtained through the additional studies is related to this particular research. Please note that, since the existence of this study was unbeknownst to the writer up until finishing the major part of this research, the proceedings in this research were done completely independent of the ones in the studies to which this appendix references.

Schweitzer and Green (2007) proposed a method in which subjects are seated in a driving simulator and watched video clips of several different driving scenes, after which they indicate on a rating scale their perceived mental demand. The video clips in the experiment are taken from the Automotive Collision Avoidance System field operational test (ACAS FOT) project, which was a naturalistic driving project in which instrumented cars were lent out for a period of 4 weeks to a total of 96 participants. Beside camera images the project also provided information on 400 engineering variables (e.g. speed, number of vehicles ahead) and other information which may be useful for relating the situations to driver workload.

Three road classes were studied: expressways, rural roads and urban roads. An independent variable in the research was the Level of Service (LOS), which is a measure used in the management of civil infrastructure which relates traffic flow with corresponding driving conditions. Three LOS were examined, which were the A, C and E level. A is determined as free flow, where traffic can easily drive at posted speed limit and there is complete mobility between lanes. C is stable flow, which has a restricted ability to maneuver through lanes but speed is still at or near speed limit. E is unstable flow, which means the road is operating at capacity. Flow is irregular and speed varies rapidly because of the lack of usable gaps to maneuver in the traffic stream, and speed is rarely at posted limit. The use of LOS in the experiment is very similar to the use of the traffic density variable in this research. The reason why the use of a categorization of the density was preferred rather than LOS is because it is very difficult to determine LOS from naturalistic data and since the density is separated into only two levels (low or high) a precise calculation of the LOS was deemed unnecessary. That being said, the low density level in this research comes closest to the A level, with the high density level coming closer to the D (approaching unstable flow) or E level.

In the 2007 research only 3 lane carriageways were considered, however a factor was included for the position of the vehicle (left, middle or right lane), whereas in this research this was not taken into account. Furthermore a factor for merging traffic was included, which was not applied on the A level of service situations since that would place them at a B level. To test the repeatability of the method, the participants where shown each clips twice. Furthermore, like in this research, each situation (combination of variables) was represented twice. Clips were counterbalanced by giving each participant a different starting clip and clip sequence.

The experiment was set up so that 5 video clips were displayed to the participant at a time. This included two anchors clips on a separate screen, which continuously looped two anchor clips of which one was a low workload situation (LOS A) and the other a high workload situations (LOS E). Next to the clips an indication of the workload level was given, which were a 2 and 6 for the low and high workload clips respectively. The other screen shows 3 video images of driving situations similar in lane width, shoulder width and curvature, with the difference between the three the level of service (LOS A for top clip, LOS C for middle clip and LOS E for bottom clip). The subjects were told to rate the demand of driving on the road on a scale from 1-10, as well as state how safe they feel it would be to (1) manually tune the radio, (2) manually dial a phone number and (3) enter a navigation destination. Before the participation in the experiment, the participants were first made familiar with the secondary tasks by performing them in the driving simulator both during driving and while not driving. The participants reported on a large amount of videos over a period of two times 30 minutes separated by a 5 minute break. Like in this experiment, the videos had a duration of 15 seconds and showed only the front view of the driver. Nighttime videos were excluded due to the video images being in black-and-white, which resulted in the inability to distinguish head- and taillights.

The use of the anchors as a reference scale during the experiment is the main difference between this research and mine. In this research, a reference frame is build using the introduction, in which videos are shown and their difficulty is related to the anchors of the RSME. While during the introduction an example is given for the range of the RSME which can be used, the participant still tends to use their own preferred range. This is likely due to the fact that the RSME uses anchors which are of an subjective nature (e.g. little effort, considerable effort, extreme effort), leaving the interpretation of the anchors open to the participant. The 2007 research assumes much greater control by allowing the participant to relate the video clips to the anchors continuously while using anchors which are more objective and are closely related to the objects which are studied. Figure F.1 shows the experimental setup used in Schweitzer and Green (2007).

Figure G.1 Experimental setup in Schweitzer and Green (2007)

Similar to this research, the 2007 research used three age categories, which were only slightly different from mine (18-30, 35-55 and 65+) with each category containing 8 subjects. The difference in sample size choice is likely a result from the much longer time period in which the experiment is taken (estimated 136 minutes). In the study the participants practiced the secondary tasks (which were not included in this research) in the simulator prior to the experiment, and judged a much larger set of situations, with each situations appearing four times.

The results of the experiment showed significant main effects for LOS, lane positions, age group as well as interaction effects for Age x Gender, LOS x age and Lane position x LOS. When comparing the results found to the results in this research, LOS can be substituted for traffic density, since the density is the most important aspect in determining LOS. A comparison shows, where applicable, the exact same effects as in this research, which are main effect for density and age, as well as interaction effects for Age x Gender and Density x age. The only difference in the findings however is that in the situations with lower density, the young group found their mean self-reported workload to be lower than the other two age categories, whereas in this research the younger groups self-reported mean ratings are always higher than the two other classes. This is likely a result of cultural differences as well as the fact that they actively avoided using students in their research, whereas in this research the younger group consisted of exclusively university students. For the Age x Gender interaction their results show a similar trend, with the elderly female participants indicating a lower workload, whereas the elderly male participants indicate a higher workload. According to Schweitzer and Green this is a common finding in many human factors studies, since older men tend to be in poorer health than women of their age, resulting in higher ratings because they have more difficulty driving.

The repeatability of the method was tested by including each video clip twice as well as have two clips of each situations (combination of variables). While the identical clips resulted in very close ratings (mean difference of 0.2 to 0.3), the mean differences for the same combinations were slightly larger (0.3 to 0.9).

Two follow up studies have been performed which relate the findings with the proposed method to data from simulator studies. In Green et al. (2011) subjects drove expressway scenarios in a driving simulator which were replicated from the naturalistic dataset. The mean workload ratings showed a high correlation with the ratings from the initial study (r=0.97), though the ratings given were lower for the video clips when compared to the simulator ratings.

Lin et al. (2012) performed another simulator study, however in this instance the focus was more on improving the anchors which were used. To test the repeatability the video clips were again rated twice, with the second time presented in a different order. The results showed a slight but statistically significant difference, however the correlation between the means of the 2 sets was 0.98, meaning the ratings were highly repeatable. Similar to the previous research, the simulator results obtained higher workload ratings than the video fragments, however with a strong correlation (r=0.84).

Appendix G.1 Implications for this research

As can be read in the previous paragraph, the two experimental designs show great similarity in concept. The main resemblance is having the participant view short video fragment rather than drive

a vehicle themselves, before filling in a rating on a subjective scale. The main difference is in the way the anchors are used, in this research the anchors leave much more room for personal interpretation, whereas in the other research the anchors used were situations very similar to the situations studied in the research. This results in that the findings in the validation study cannot necessarily be applied to this study. Due to the similarity between the two research methods it can be assumed that a similar validity can be found in this study, however it is likely that the repeatability is lower as a result of the difference in anchors used. The findings of the study do however show promise, seeing the same main and interaction effects are found between both studies. Further research towards the subject could study the relation between ratings found with the video fragments method and real life driving, rather than simulated driving.

Appendix H – Full results rANOVA

Table H.1 - Full results rANOVA motor way situations

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Density	Pillai's Trace	,749	169,815[b]	1,000	57,000	,000	,749
	Wilks' Lambda	,251	169,815[b]	1,000	57,000	,000	,749
	Hotelling's Trace	2,979	169,815[b]	1,000	57,000	,000	,749
	Roy's Largest Root	2,979	169,815[b]	1,000	57,000	,000	,749
Density * Age_cat	Pillai's Trace	,121	3,930[b]	2,000	57,000	,025	,121
	Wilks' Lambda	,879	3,930[b]	2,000	57,000	,025	,121
	Hotelling's Trace	,138	3,930[b]	2,000	57,000	,025	,121
	Roy's Largest Root	,138	3,930[b]	2,000	57,000	,025	,121
Lanes	Pillai's Trace	,006	,370[b]	1,000	57,000	,545	,006
	Wilks' Lambda	,994	,370[b]	1,000	57,000	,545	,006
	Hotelling's Trace	,006	,370[b]	1,000	57,000	,545	,006
	Roy's Largest Root	,006	,370[b]	1,000	57,000	,545	,006
Lanes * Age_cat	Pillai's Trace	,142	4,703[b]	2,000	57,000	,013	,142
	Wilks' Lambda	,858	4,703[b]	2,000	57,000	,013	,142
	Hotelling's Trace	,165	4,703[b]	2,000	57,000	,013	,142
	Roy's Largest Root	,165	4,703[b]	2,000	57,000	,013	,142
HGV	Pillai's Trace	,465	49,541[b]	1,000	57,000	,000	,465
	Wilks' Lambda	,535	49,541[b]	1,000	57,000	,000	,465
	Hotelling's Trace	,869	49,541[b]	1,000	57,000	,000	,465
	Roy's Largest Root	,869	49,541[b]	1,000	57,000	,000	,465
HGV * Age_cat	Pillai's Trace	,065	1,987[b]	2,000	57,000	,147	,065
	Wilks' Lambda	,935	1,987[b]	2,000	57,000	,147	,065
	Hotelling's Trace	,070	1,987[b]	2,000	57,000	,147	,065
	Roy's Largest Root	,070	1,987[b]	2,000	57,000	,147	,065
Density * Lanes	Pillai's Trace	,000	,009[b]	1,000	57,000	,925	,000
	Wilks' Lambda	1,000	,009[b]	1,000	57,000	,925	,000
	Hotelling's Trace	,000	,009[b]	1,000	57,000	,925	,000
	Roy's Largest Root	,000	,009[b]	1,000	57,000	,925	,000
Density * Lanes * Age_cat	Pillai's Trace	,069	2,103[b]	2,000	57,000	,131	,069
	Wilks' Lambda	,931	2,103[b]	2,000	57,000	,131	,069
	Hotelling's Trace	,074	2,103[b]	2,000	57,000	,131	,069
	Roy's Largest Root	,074	2,103[b]	2,000	57,000	,131	,069
Density * HGV	Pillai's Trace	,162	11,009[b]	1,000	57,000	,002	,162
	Wilks' Lambda	,838	11,009[b]	1,000	57,000	,002	,162
	Hotelling's Trace	,193	11,009[b]	1,000	57,000	,002	,162
	Roy's Largest Root	,193	11,009[b]	1,000	57,000	,002	,162
Density * HGV * Age_cat	Pillai's Trace	,071	2,193[b]	2,000	57,000	,121	,071
	Wilks' Lambda	,929	2,193[b]	2,000	57,000	,121	,071
	Hotelling's Trace	,077	2,193[b]	2,000	57,000	,121	,071
	Roy's Largest Root	,077	2,193[b]	2,000	57,000	,121	,071
Lanes * HGV	Pillai's Trace	,033	1,939[b]	1,000	57,000	,169	,033
	Wilks' Lambda	,967	1,939[b]	1,000	57,000	,169	,033
	Hotelling's Trace	,034	1,939[b]	1,000	57,000	,169	,033
	Roy's Largest Root	,034	1,939[b]	1,000	57,000	,169	,033
Lanes * HGV * Age_cat	Pillai's Trace	,040	1,185[b]	2,000	57,000	,313	,040
	Wilks' Lambda	,960	1,185[b]	2,000	57,000	,313	,040
	Hotelling's Trace	,042	1,185[b]	2,000	57,000	,313	,040
	Roy's Largest Root	,042	1,185[b]	2,000	57,000	,313	,040
Density * Lanes * HGV	Pillai's Trace	,001	,034[b]	1,000	57,000	,855	,001
	Wilks' Lambda	,999	,034[b]	1,000	57,000	,855	,001
	Hotelling's Trace	,001	,034[b]	1,000	57,000	,855	,001
	Roy's Largest Root	,001	,034[b]	1,000	57,000	,855	,001
Density * Lanes * HGV * Age_cat	Pillai's Trace	,035	1,023[b]	2,000	57,000	,366	,035
	Wilks' Lambda	,965	1,023[b]	2,000	57,000	,366	,035
	Hotelling's Trace	,036	1,023[b]	2,000	57,000	,366	,035
	Roy's Largest Root	,036	1,023[b]	2,000	57,000	,366	,035

Table H.2 - Full rANOVA results VRU

Within Subjects Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
VRU	Pillai's Trace	,732	36,824[c]	4,000	54,000	,000	,732
	Wilks' Lambda	,268	36,824[c]	4,000	54,000	,000	,732
	Hotelling's Trace	2,728	36,824[c]	4,000	54,000	,000	,732
	Roy's Largest Root	2,728	36,824[c]	4,000	54,000	,000	,732
VRU * Age_cat	Pillai's Trace	,432	3,788	8,000	110,000	,001	,216
	Wilks' Lambda	,577	4,265[c]	8,000	108,000	,000	,240
	Hotelling's Trace	,715	4,740	8,000	106,000	,000	,263
	Roy's Largest Root	,692	9,514[d]	4,000	55,000	,000	,409

Table H.3 - Full rANOVA results Weather

Within Subjects Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Weather	Pillai's Trace	,680	28,725[c]	4,000	54,000	,000	,680
	Wilks' Lambda	,320	28,725[c]	4,000	54,000	,000	,680
	Hotelling's Trace	2,128	28,725[c]	4,000	54,000	,000	,680
	Roy's Largest Root	2,128	28,725[c]	4,000	54,000	,000	,680
Weather * Age_cat	Pillai's Trace	,130	,955	8,000	110,000	,475	,065
	Wilks' Lambda	,874	,940[c]	8,000	108,000	,487	,065
	Hotelling's Trace	,140	,926	8,000	106,000	,498	,065
	Roy's Largest Root	,092	1,265[d]	4,000	55,000	,295	,084