# INFORMATION SOCIETY TECHNOLOGIES (IST) PROGRAMME

Information Society Technologies

## AIDE
## IST-1-507674-IP

# Subjective Assessment methods for Workload

| | | | |
|---|---|---|---|
| Deliverable No. | | **D2.2.6** | |
| SubProject No. | **SP2** | SubProject Title | **Evaluation and assessment methodology** |
| Workpackage No. | **WP2.2** | Workpackage Title | **Driver workload and distraction assessment methods and tools** |
| Activity No. | **2.2.6** | Activity Title | **Subjective workload assessment methods** |
| Authors | | **E. Chin and F. Nathan (PSA), A. Pauzié and J. Manzano (INRETS), E. Nodari, C. Cherri , A. Rambaldini , A. Toffetti (CRF), M. Marchitto (KITE)** | |
| Status | | **Final** | |
| File Name: | | **AIDE_SP2_WP2.2_D226.doc** | |
| Project start date and duration | | **01 March 2004, 48 Months** | |

# List of abbreviations

BMDMW : Behavioural Markers of Driver Mental Workload
DALI : Driving Activity Load Index
NASA-TLX : NASA Task Load indeX
PSA-TLX : PSA Task Load indeX
ACC : Autonomous Cruise Control
FCWS : Frontal Collision Warning System
HCS : High context + system
HC : high context
LCS : low context + system
LC : low context

# Table of content

List of abbreviations ..... 2
List of Figures ..... 5
List of Tables ..... 9
EXECUTIVE SUMMARY ..... 11
1 Introduction ..... 13
2 Common plan/methodology/theoretical framework : state of the art (KITE, CRF, PSA, INRETS) ..... 17
2.1 "OFF-LINE" ASSESSMENT METHODS ..... 18
2.1.1 NASA Task Load Index (TLX) ..... 18
2.1.2 SUBJECTIVE WORKLOAD DOMINANCE (SWORD) ..... 20
2.1.3 SUBJECTIVE WORKLOAD ASSESSMENT TECHNIQUE (SWAT) ..... 21
2.1.4 COOPER-HARPER SCALE and MODIFIED COOPER-HARPER SCALE (MCH). 22
2.1.5 MALVERN CAPACITY ESTIMATE (MACE) ..... 25
2.1.6 ANALYTICAL HIERARCHY PROCESS (AHP) ..... 25
2.1.7 CREW STATUS SURVEY ..... 26
2.1.8 FLIGHT WORKLOAD QUESTIONNAIRE ..... 27
2.1.9 HONEYWELL COOPER-HARPER RATING SCALE ..... 28
2.1.10 NASA BIPOLAR RATING SCALE ..... 30
2.1.11 EQUAL-APPEARING INTERVALS ..... 31
2.1.12 MAGNITUDE ESTIMATION ..... 31
2.2 "ON-LINE" ASSESSMENT TOOLS ..... 32
2.2.1 DEFENCE RESEARCH AGENCY WORKLOAD SCALE (DRAWS) ..... 32
2.2.2 INSTANTANEOUS SELF ASSESSMENT (ISA) ..... 32
2.2.3 BEDFORD SCALE ..... 33
2.2.4 DYNAMIC WORKLOAD SCALE ..... 34
2.2.5 HART AND HAUSER RATING SCALE ..... 35
2.3 WORKLOAD PREDICTION TOOLS ..... 36
2.3.1 PERFORMANCE AND USABILITY MODELLING IN ATM (PUMA) ..... 36
2.3.2 WINCREW ..... 36
2.4 THE THREE TOOLS SELECTED FOR FIELD TESTS ..... 37
2.4.1 BMDMW ..... 37
2.4.2 DALI ..... 40
2.4.3 PSA-TLX ..... 46
2.5 CONCLUSION of state of the art ..... 56
3 Input from previous AIDE work ..... 60
4 Contribution to overall AIDE objectives ..... 61
5 Experiment 1 (CRF) ..... 62
5.1 Method ..... 62
5.1.1 Intended users ..... 62
5.1.2 Participants ..... 62
5.1.3 Apparatus ..... 63
5.1.4 Data analysis ..... 67
5.1.6 Tasks/Systems (e.g. IVIS tasks or ADAS systems, both…) ..... 74
5.1.7 Design and Procedure ..... 75
5.2 Results ..... 76
5.2.1 PSA-TLX results ..... 76
5.2.2 DALI results ..... 83
5.2.3 BMDMW results ..... 85

5.2.4 Objective results ... 86
5.2.5 Comparison between Objective and Subjective results ... 95
5.3 Discussion on CRF results ... 103
6 Experiment 2 (INRETS) ... 104
6.1 Method ... 104
6.1.1 Participants ... 105
6.1.2 Apparatus ... 105
6.1.3 Data analysis ... 107
6.1.4 Comparison of conditions ... 107
6.1.5 Tasks/Systems (e.g. IVIS tasks or ADAS systems, both…) ... 108
6.1.6 Design and Procedure ... 109
6.1.7 Inferential statistical measures ... 109
6.2 Results ... 110
6.2.1 DALI ... 110
6.2.2 PSA-TLX results ... 117
6.2.3 Analysis of the results from BMDMW ... 133
6.2.4 Synthesis of final evaluation of questionnaires ... 139
6.3 Discussion on INRETS results ... 140
7 Experiment 3 (PSA) ... 142
7.1 Method ... 143
7.1.1 Participants ... 143
7.1.2 Apparatus ... 143
7.1.3 Data analysis ... 144
7.1.4 Inferential statistical measures ... 144
7.1.5 Comparison of conditions ... 144
7.1.6 Tasks/Systems (e.g. IVIS tasks or ADAS systems, both…) ... 145
7.1.7 Design and Procedure ... 148
7.2 Results ... 149
7.2.1 PSA-TLX ... 149
7.2.2 DALI ... 167
7.2.3 Evaluation by drivers: final questionnaire ... 170
7.3 Discussion on PSA results ... 171
8 Comparison between different Methods/metrics ... 172
8.1 Description of procedure ... 172
8.2 Criteria for comparison ... 172
8.3 Assessment of the considered (single) methods ... 173
8.4 Experimental evaluation of some selected methods ... 181
9 Conclusions: Specification of suggested method/metric to be used in the AIDE test regime ... 183
10 Innovation ... 188
11 References ... 189
12 Appendix ... 192

# List of Figures

Fig. 1-1 : Workload evaluation of AIDE prototype ..... 13
Fig. 1-2 : Schema of control process as effect of activity, adapted from Leplat and Cuny, 1977 to driving ..... 14
Fig. 1-3 : Overview of the three workload subjective methods proposed for AIDE system evaluation ..... 14
Fig. 2-1: Graphic example of the composition of a weighted workload score. ..... 19
Fig. 2-2 : An example of a SWORD evaluation form: PA, UA, etc…are the particular tasks compared each other. (reproduced from Vidulich et al, 1991). ..... 20
Fig. 2-3 : an example of a SWORD judgment matrix (reproduced from Vidulich et al., 1991) ..... 20
Fig. 2-4 : The decision tree of the Cooper-Harper Scale ..... 23
Fig. 2-5 : The Modified Cooper-Harper Rating Scale (Wierwille and Casali 1983) ..... 24
Fig. 2-6 : AHP rating scale ..... 25
Fig. 2-7 : Honeywell Cooper-Harper Rating Scale ..... 29
Fig. 2-8 : The hierarchical decision tree of Bedford Scale ..... 34
Fig. 2-9: Dynamic Workload Scale ..... 34
Fig. 2-10 : Hart and Hauser rating scale ..... 35
Fig. 2-11 : The graphic interface of WinCrew. ..... 36
Fig. 2-12: DALI-Evaluation of a Guidance/Navigation system with the DALI questionnaire 42
Fig. 2-13: DALI-Evaluation of a Mobile Phone System with the DALI questionnaire ..... 43
Fig-2-14 : Overview of the DALI ‘s tool to evaluate the various components of the driving task workload ..... 44
Fig. 2-15 : De Waard Model of mental workload (1996) ..... 48
Fig. 2-16 : PSA-TLX scale (example of lateral control evaluation) ..... 50
Fig. 2-17 : PSA-TLX - Performance safety in driving sub-tasks ..... 52
Fig. 2-18 : PSA-TLX - Effort in driving sub-tasks ..... 52
Fig. 2-19 : PSA-TLX - Frequency of different types of compromise (night vision system) ... 53
Fig. 2-20 : PSA-TLX - Frequency of different types of compromise per driver (night vision system) ..... 53
Fig. 2-21: PSA-TLX - Frequency of types of compromises per situations compared (IVIS) . 55
Fig. 5-1: The CRF Driving simulator architecture ..... 63
Fig. 5-2: Examples of scenarios populated by vehicular traffic ..... 64
Fig. 5-3: Simulated Scenario, shrinkage with static signs ..... 65
Fig. 5-4: Simulated Scenario, another view of shrinkage ..... 65
Fig. 5-5: Types of Segments in VR Trial ..... 66
Fig. 5-6: General scheme of the three experimental conditions ..... 68
Fig. 5-7: Global Mean Speed Dispersion ..... 70
Fig. 5-8: Square Errors Distribution ..... 71
Fig. 5-9: Square Errors Normalized Distribution ..... 71
Fig. 5-10: Relationship among Distributions ..... 72
Fig. 5-11: Comparison between Speed Distribution and Normalized Speed Dispersion ..... 73
Fig. 5-12: Segments taken into account in specific secondary task analysis ..... 74
Fig. 5-13: Difficult Task Scheme with IVIS ..... 75
Fig. 5-14: PSA-TLX - Differences between VR Experts and VR Non-Experts ..... 77
Fig. 5-15: PSA-TLX - Overall Driving - Performance ..... 78
Fig. 5-16: PSA-TLX - Overall Driving – Effort ..... 78
Fig. 5-17: PSA-TLX – Single Sub-Tasks Score – Performance ..... 79
Fig. 5-18: PSA-TLX – Performance components ..... 80
Fig. 5-19: PSA-TLX – Single Sub-Tasks Score – Effort ..... 80

Fig. 5-20: PSA-TLX – Mean distribution of Effort over Subtasks ..... 81
Fig. 5-21: PSA-TLX – Graphical representation of Compromise levels ..... 82
Fig. 5-22: DALI - Difference among Conditions respect to different Factors and Global Workload and summarizing table with Wilcoxon Text ..... 84
Fig. 5-23: BMDMW – Global Demand across Conditions ..... 85
Fig. 5-24: BMDMW – Factors of Demand across Conditions ..... 85
Fig. 5-25: Global Index of Conditions ..... 87
Fig. 5-26: Objective Results - Synthetical Variability Indices for Brake segment ..... 88
Fig. 5-27: Objective Results – Variability of Brake reaction time and Warning duration in Brake segment – Medium Visibility ..... 89
Fig. 5-28: Objective Results - Variability of Brake reaction times and Warning duration in Brake segment – Good Visibility ..... 89
Fig. 5-29: Objective Results - Synthetical Variability Indices for Shrinkage segment ..... 90
Fig. 5-30: Objective Results – Relative distances variability for Shrinkage segment ..... 91
Fig. 5-31: Objective Results – Synthetical Variability Indices for Obstacle segment ..... 92
Fig. 5-32: Deeper analysis of secondary task impact – Lateral displacement variations ..... 94
Fig. 5-33: Deeper analysis of secondary task impact – Mean speed variations ..... 94
Fig. 5-34: Comparison between PSA-TLX results and Objective Results - Effort ..... 95
Fig. 5-35: Comparison between PSA-TLX results and Objective Results - Effort ..... 96
Fig. 5-36: Comparison between PSA-TLX results and Objective Results - Performance ..... 97
Fig. 5-37: Comparison between DALI Visual Demand and Objective Results ..... 98
Fig. 5-38: Comparison between DALI Interference and Objective Results ..... 99
Fig. 5-39: Comparison between DALI Stress and Objective Results ..... 99
Fig. 5-40: Comparison between DALI Global Demand and Objective Results ..... 100
Fig. 5-41: Comparison between BMDMW Results and Objective Results ..... 101
Fig. 6-1: The in-vehicle equipments of MARGO ..... 105
Fig. 6-2 : Guidance system used for INRETS experiment ..... 106
Fig. 6-3: Screen used for INRETS experiment ..... 106
Fig. 6-4: Guidance system and screen integrated in the test vehicle ..... 106
Fig. 6-5: Computer used to send information to the front screen ..... 106
Fig. 6-6: Computer recording parameters of the vehicle ..... 107
Fig. 6-7: Turn over between the sessions ..... 108
Fig. 6-8: DALI -Values for each factor and for global score for the 4 experimental sessions ..... 110
Fig. 6-9: DALI -Values for each weighted factor and for global score for the 4 experimental sessions ..... 114
Fig. 6-10: DALI -Comparison between young and older drivers ..... 116
Fig. 6-11: PSA-TLX -Mean global scores for performance safety ..... 117
Fig. 6-12: PSA-TLX -Mean global scores for effort felt ..... 119
Fig. 6-13: PSA-TLX -Compromise for global performance and global effort ..... 122
Fig. 6-14: PSA-TLX -mean global scores for global performance and effort felt ..... 122
Fig. 6-15: PSA-TLX -mean scores of performance safety for each driving sub-task ..... 123
Fig. 6-16: PSA-TLX -Mean scores of effort felt for each driving sub-task ..... 125
Fig. 6-17: PSA-TLX -Management of performance ..... 126
Fig. 6-18: PSA-TLX -Managment of effort ..... 127
Fig. 6-19: PSA-TLX -Compromise between performance and effort in control of the lateral motion ..... 128
Fig. 6-20: PSA-TLX -Compromise between global performance and global effort for LCS and HC conditions ..... 129
Fig. 6-21 : BMDMW-Variation of global workload across conditions ..... 133

Fig. 6-22: BMDMW-Scores got for the questions 1, 4, 8, 10 and 15 ..... 134
Fig. 6-23: BMDMW-Scores got for the questions 16, 17, 18, 21, 25 and 26 ..... 134
Fig. 6-24: BMDMW-Scores got for the questions 27, 28, 29, 30, 32 and 33 ..... 135
Fig. 6-25: BMDMW-Scores got for the questions 2, 3, 5, 6 and 7 ..... 136
Fig. 6-26: BMDMW-Scores got for the questions 9, 11, 12, 13 and 14Figure 53..... 136
Fig. 6-27: BMDMW-Scores got for the questions 19, 20, 22, 23, 24 and 31 ..... 137
Fig. 7-1 : Test-vehicle ..... 143
Fig. 7-2 : Itinerary ..... 145
Fig. 7-3 : ACC system HMI ..... 146
Fig. 7-4 : Manual controls of ACC system ..... 147
Fig. 7-5 : PSA-TLX - Driving performance – mean of global scores..... 149
Fig. 7-6 : PSA-TLX - Driving performance –individual global scores ..... 150
Fig. 7-7 : PSA-TLX - Performance to sub-tasks – group's trend – mean of individual scores ..... 151
Fig. 7-8 : PSA-TLX -Performance in lateral control – individual scores ..... 152
Fig. 7-9 : PSA-TLX -Performance in longitudinal control – individual scores..... 153
Fig. 7-10 : PSA-TLX - Performance in reacting to dynamic environment – individual scores ..... 153
Fig. 7-11 : PSA-TLX - Performance in reacting to static environment – individual scores.. 154
Fig. 7-12 : PSA-TLX -Performance management ..... 155
Fig. 7-13 : PSA-TLX -Driving effort - Mean of global scores ..... 156
Fig. 7-14 : PSA-TLX -Driving effort – Individual global scores..... 156
Fig. 7-15 : PSA-TLX -Driving effort per driving sub-tasks – group's trend – mean of individual scores..... 157
Fig. 7-16 : PSA-TLX -Effort for lateral control – individual scores..... 158
Fig. 7-17 : PSA-TLX -Effort for longitudinal control – individual scores ..... 159
Fig. 7-18 : PSA-TLX -Effort for reacting to dynamic environment – individual scores..... 159
Fig. 7-19 : PSA-TLX -Effort for reacting to static environment – individual scores ..... 160
Fig. 7-20 : PSA-TLX -Effort management ..... 162
Fig. 7-21 : PSA-TLX -Changes in compromises efficiency among situations..... 163
Fig. 7-22: PSA-TLX -Frequency of different types of compromise Effort/Performance per driver – Driving with ACC (3) vs. driving without ACC (1) - learning vehicle..... 164
Fig. 7-23: PSA-TLX -Frequency of different types of compromise Effort/Performance per driver – Driving with ACC (3) vs. driving without ACC (1) - learning vehicle..... 164
Fig. 7-24: PSA-TLX -Frequency of different types of compromise Effort/Performance per driver – Driving without ACC (4) vs. driving without ACC (1) - learning vehicle..... 164
Fig. 7-25: PSA-TLX -Types of compromise Effort/Performance per sub-task – Driving with ACC vs. driving without ACC (learning vehicle)..... 165
Fig. 7-26: PSA-TLX -Types of compromise Effort/Performance per sub-task – Driving without ACC vs. driving with ACC ..... 165
Fig. 7-27 : PSA-TLX -Types of compromise Effort/Performance per driver – Driving without ACC vs. driving without ACC (learning vehicle)..... 165
Fig. 7-28 : DALI - Mean raw scores ..... 167
Fig. 7-29 : DALI - Mean Weighted scores..... 168
Fig. 8-1 : Questionnaire presentation – all experiments..... 173
Fig. 8-2 : Duration of questionnaire filling in – all experiments..... 174
Fig. 8-3 : Score's choice – all experiments ..... 174
Fig. 8-4 : Questions understanding – all experiments ..... 175
Fig. 8-5 : Concept understanding – all experiments..... 175
Fig. 8-6 : understanding of written explanations – all experiments ..... 176

Fig. 8-7 : Explanations of experimenter (necessary)– all experiments ..... 176
Fig. 8-8 : Explanations of experimenter (sufficient) – all experiments ..... 177
Fig. 8-9 : Required effort to fill in questionnaires – all experiments ..... 177
Fig. 8-10 : Evolution of effort required to fill in questionnaire – all experiments ..... 178
Fig. 8-11 : Exhaustivity of questionnaire – all experiments ..... 178
Fig. 8-12 : Effort required to fill-in questionnaires – all experiments ..... 179
Fig. 8-13 : level of fatigue at the end of experiment – All experiments ..... 179
Fig. 8-14 : Reasons of tiredness – all experiments ..... 180

# List of Tables

Table 1-1: Experimental design of all experiments ..... 15
Table 1-2: Comparisons of conditions ..... 16
Table 2-1: Definitions of the six workload subscales of NASA TLX. ..... 18
Table 2-2 : The dimensions and scales of SWAT (reproduced from Wickens et al., 2000) .... 21
Table 2-3 : Definitions of AHP scale descriptors. ..... 26
Table 2-4 : Crew Status Survey. ..... 27
Table 2-5 : NASA Bipolar Rating Scale Definitions ..... 30
Table 2-6: NASA Bipolar Rating Scale ..... 31
Table 2-7 Description of DALI's factors ..... 41
Table 2-8 : DALI data processing ..... 45
Table 2-9 - Correspondence between Schlegel's model and PSA-TLX driving task components ..... 47
Table 2-10 : PSA-TLX components of workload ..... 49
Table 2-11 - PSA-TLX data processing ..... 51
Table 2-12 : PSA-TLX - ANOVA on driving performance ..... 54
Table 2-13 : PSA-TLX - ANOVA on driving effort ..... 54
Table 2-14 : MCH Scale evaluation ..... 57
Table 2-15 : SWAT evaluation ..... 57
Table 2-16 : NASA-TLX evaluation ..... 58
Table 5-1: Summary of Metrics in Objective Data Analysis ..... 69
Table 5-2: Experiment Phases and timings ..... 76
Table 5-3: PSA-TLX – Significances among pair comparison of Conditions for Performance ..... 79
Table 5-4: PSA-TLX – Significances among pair comparison of Conditions for Effort ..... 80
Table 5-5: PSA-TLX – Compromise Effort-Performance; Summary of Distributions ..... 81
Table 5-6: PSA-TLX – Rating of Compromise Quality among Sub-tasks ..... 82
Table 5-7: DALI - Relationship between Weighted and Raw Data ..... 83
Table 5-8: DALI – Significances among DALI factors for different conditions ..... 84
Table 5-9: BMDMW – Significances among BMDMW factors for different conditions ..... 86
Table 5-10: Objective Results – significances in Brake Segment ..... 88
Table 5-11: Objective Results - Analytical Significances in Shrinkage Segment ..... 91
Table 5-12: Objective Results – Lateral displacement significances in Task Segment ..... 94
Table 5-13: Objective Results – Mean speed significances in Task Segment ..... 95
Table 6-1 - DALI -Non parametric Wilcoxon test used to analyse ..... 110
Table 6-2 - DALI -Non parametric Wilcoxon test used to analyse weighted factors ..... 114
Table 6-3 - DALI -Spearman test to analyse the correlation between factors and weighted factors ..... 115
Table 6-4 - PSA-TLX -Wilcoxon test used to compare global scores performance ..... 117
Table 6-5 - PSA-TLX -Analysis of mean values for each experimental session for performance safety ..... 118
Table 6-6 - PSA-TLX -Wilcoxon test used to compare global scores effort ..... 119
Table 6-7 : PSA-TLX -Analysis of mean values for each experimental sessions for effort felt ..... 119
Table 6-8 - PSA-TLX -Wilcoxon test used to compare experimental sessions for performance for each sub-task ..... 123
Table 6-9 - PSA-TLX -Analysis of mean values for each experimental session for performance got for each sub-task ..... 124

Table 6-10 - PSA-TLX -Wilcoxon test used to compare experimental sessions for effort for each sub-task ..... 125
Table 6-11 - PSA-TLX -Analysis of mean values for each experimental session for effort for each sub-task ..... 125
Table 6-12 - PSA-TLX -Analyse of global compromise Performance/Effort for LCS and HC conditions ..... 129
Table 6-13 - PSA-TLX -Summary of global compromise Performance/Effort for LCS and HC conditions ..... 129
Table 6-14 : BMDMW-Wilcoxon test used to compare global workload for the experimental sessions ..... 133
Table 6-15 : BMDMW-Wilcoxon test used to compare scores of direct assessment for the experimental sessions ..... 135
Table 6-16: BMDMW-Wilcoxon test used to compare scores of invers assessment for the experimental sessions ..... 137
Table 7-1 : In-vehicle devices ..... 143
Table 7-2 : Comparison of conditions ..... 144
Table 7-3 : Design procedure (minutes) ..... 148
Table 7-4 : PSA-TLX - Performance to sub-tasks - Spearman Ranks' Correlation test ..... 151
Table 7-5 : PSA-TLX - Performance to sub-tasks – group's trend – Test of Wilcoxon ..... 151
Table 7-6 : PSA-TLX -Effort to sub tasks- Spearman ranks' Correlation test– raw score ..... 157
Table 7-7 : PSA-TLX -Effort to sub-tasks – group's trend - T Wilcoxon – raw scores ..... 158
Table 7-8 : DALI - Rank correlation of Spearman – raw scores ..... 167
Table 7-9 : DALI - Wilcoxon test – raw scores ..... 168
Table 7-10 : DALI - Rank correlation of Spearman – weighted scores ..... 168
Table 7-11 : DALI - Wilcoxon test – weighted scores ..... 169
Table 7-12 : DALI - Bravais – Pearson Correlation between raw scores and weighted scores ..... 169
Table 9-1 – Recommendations for BMDMW questionnaire ..... 184
Table 9-2 – Recommendations for DALI questionnaire ..... 185
Table 9-3 – Recommendations for PSA-TLX questionnaire ..... 187

# EXECUTIVE SUMMARY

The objective of task 2.2.6. is to recommend the more suitable workload subjective method to evaluate IVIS and ADAS, and AIDE system, among existing methods.

This report presents a state of the art on subjective methods used to evaluate workload and results of three experiments the purpose of which is to evaluate the sensitivity, the advantages, the drawbacks and the limits of three existing tools (questionnaires) dedicated to evaluate several components of workload of different natures (see section 2.4 of the deliverable for more details on each tool):

- **BMDMW** questionnaire - Behavioural Markers of Driver Mental Workload – designed by University of Roma. This tool is a rating of 34 questions dealing with behavioural patterns related to driving task and driver's state (see section 2.4.1 and independent appendix I.1)
- **DALI** questionnaire - Driving Activity Load Index – designed by INRETS-LESCOT which is an adaptation of the NASA-TLX questionnaire for automobile and in particular for design of IVIS. Mental workload is evaluated in terms of task demands-visual demand, auditory demand, tactile demand, temporal demand – and in terms of effort of attention, interference and stress (see section 2.4.2 and independent appendix II.2)
- **PSA-TLX** questionnaire (PSA-Task Load index) designed by PSA-Peugeot Citroen (Research department), is a driving task-oriented subjective measure of driving workload. It is dedicated to evaluate driving workload (compromise between supplied effort and reached performance considering the driver's state) when using IVIS or when driving assisted by an ADAS, (see section 2.4.3 and independent appendix III.2).

Three experiments have been achieved with three case studies in different test environments:

- Road experiment on IVIS – *INRETS*: route guidance system and paper map (additional artificial tasks)
- Road experiment on ADAS - *PSA*: Autonomous Cruise Control (ACC)
- Simulator experiment on ADAS – *CRF*: Frontal Collision Warning System (FCWS)

The three tools have not been tested in the three experiments as it was planned at the beginning of the task, because:

- all tools were not available at the beginning of tests and tests could not be postponed considering the time plan,
- the administration of three different questionnaires may induce potential interference between concepts evaluated and the filling procedure asked to participants (each questionnaire presents specific principles),
- The administration of the three tools is time consuming and experimental protocol could not afford the administration of all three tools in each experiment.

Results showed that the three methods are sensitive to different level of task's demands and are complementary as:

- They are designed for different purposes.
- Several components of workload of different nature are evaluated.
- They present different level of sensitivity, depending on the test environment (simulator or road), the nature and level of demands (ACC, FCWS or route guidance system/paper map; driver's characteristics - experts vs. non-experts of simulator driving).

The different questionnaires are not considered as directly concurrent as they are complementary.

Consequently, the methods have been recommended depending on the purpose of the evaluation (HMI comparison, impact on driving safety, driving behavioural compensation to cope with higher demands…) and the practical constraint (e.g. time constraint to achieve experiments, experimenter training/experience…).

Tests conducted inside T226 have allowed defining the area of use and the purpose for each method, in order to specify the most suitable tool for each case study.

# 1 Introduction

The objective of task 2.2.6. is to recommend the most suitable driver workload subjective method to evaluate IVIS and ADAS, and AIDE system, among existing methods.
AIDE prototype developed in SP3 is dedicated to manage inputs from all functions directed to the driver (ADAS and IVIS) according to many parameters (driver's availability depending on driver's state, driving environment, IVIS or ADAS demands…), with a global objective of reduce driving workload by integration and adaptivity.
AIDE prototype may induce workload even if it aims to be intelligent and decrease driver's "overload" risk by avoiding conflicts between several functions or between a function and a specific driving situation.

Below the components of workload which can be measured by the three tools evaluated in T226 are described (see Fig. 1-1).

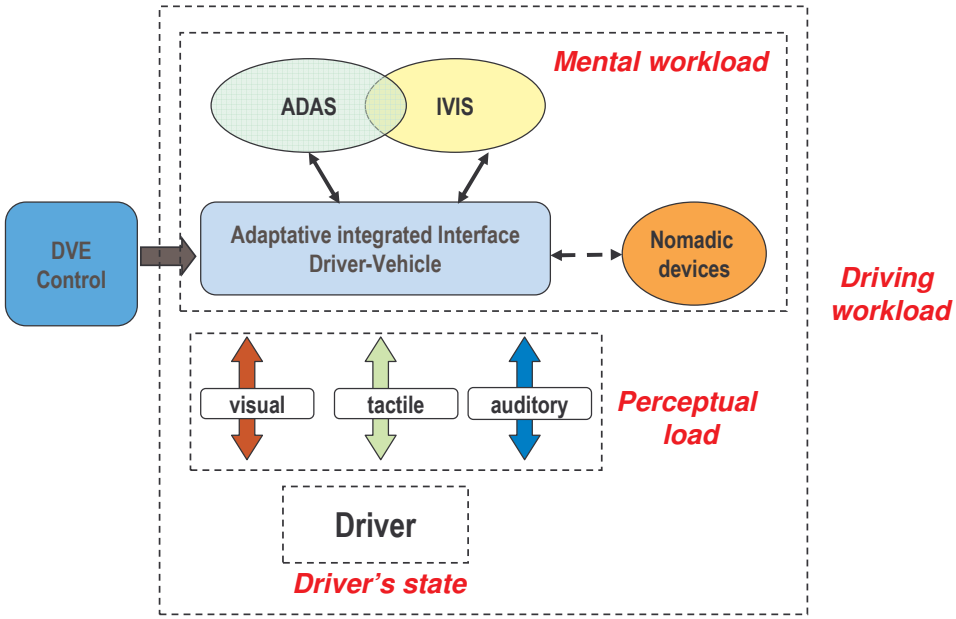



**Fig. 1-1 : Workload evaluation of AIDE prototype**

- Definition of workload

Schlegel (1993) defined driving workload as "amount of resources (or abilities) allocated by the driver, in terms of effort and attention, to achieve the driving task."
Driving workload reflects the driving task's demands on an individual driver and his ability to cope with the demands to respect an acceptable level of performance.
It depends on :

- Task (demands, complexity, difficulties)
- Driver (age, fatigue, abilities, experience…)
- Environment (dynamic / changing, familiar…)

Workload can be measured:

- Either by task's demands (temporal, cognitive, perceptual..) as it is proposed in DALI questionnaire,
- Or as compromise between task's demands and task's performance as proposed by PSA-TLX and BMDMW.

Workload induced by a driving activity results from a control process between task's demands and effect on task's performance: task's demands determine workload and performance and reached performance influences task's demands and workload (see Fig. 1-2)

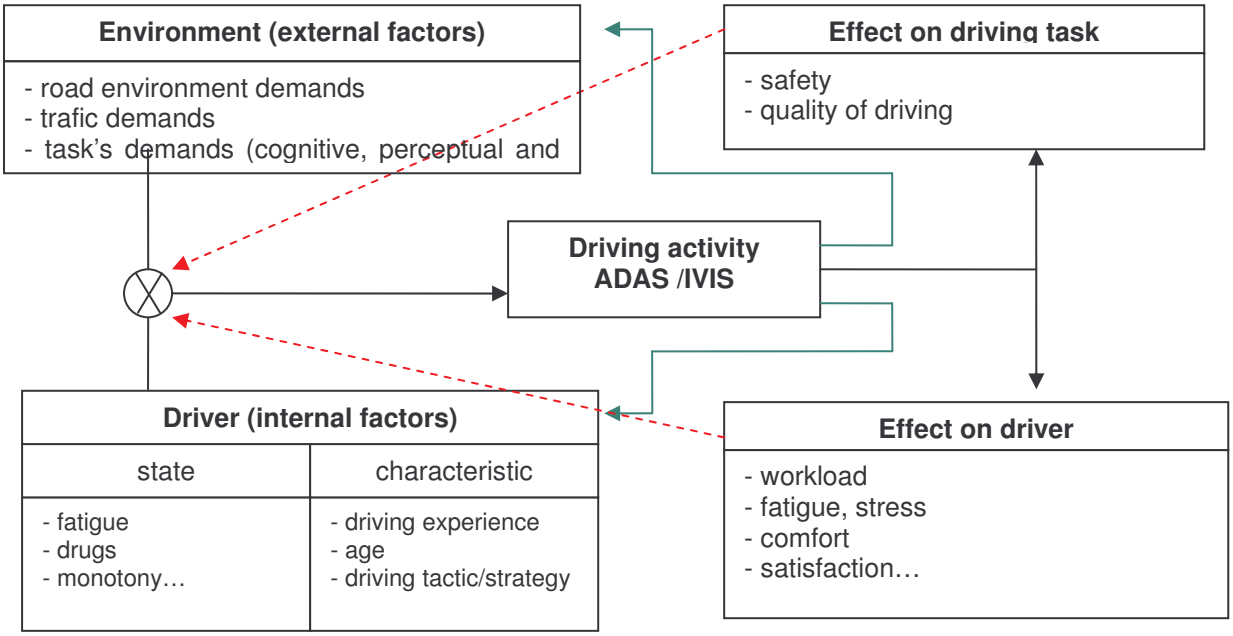



**Fig. 1-2 : Schema of control process as effect of activity, adapted from Leplat and Cuny, 1977 to driving**

Overview of 3 subjective workload assessment methods proposed for AIDE Evaluation

Below the three tools are presented depending on measured components of workload and their nature (see Fig. 1-3).

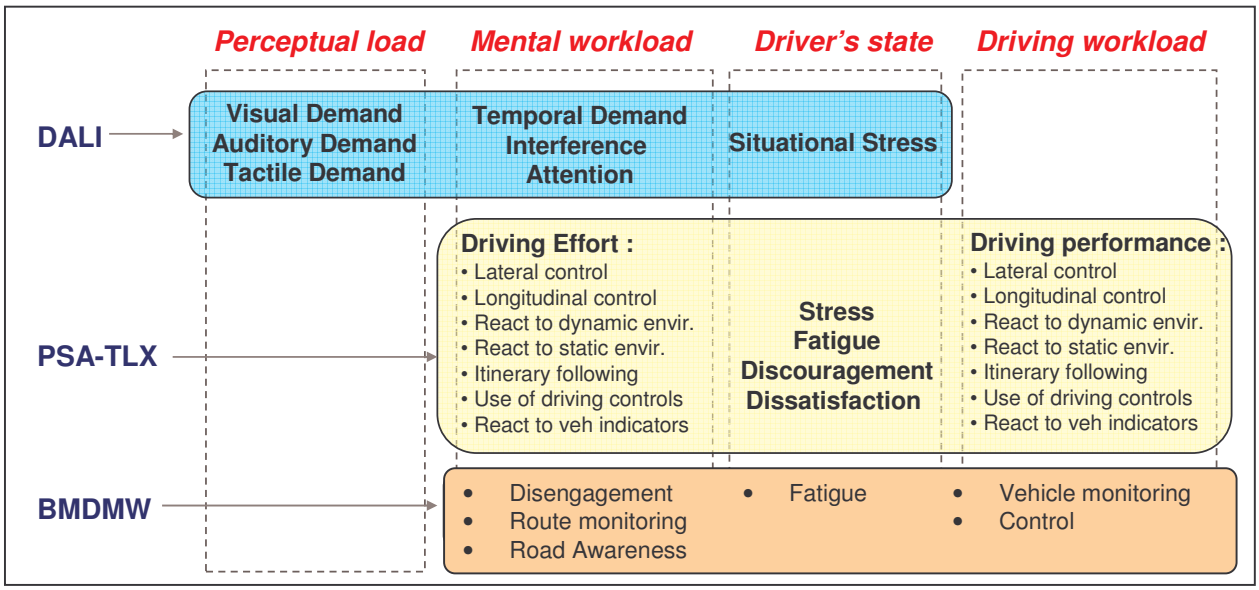



**Fig. 1-3 : Overview of the three workload subjective methods proposed for AIDE system evaluation**

The three tools are dedicated to evaluate several components of workload of different natures; they are complementary as covering different components of workload are different:

- **Perceptual load** is evaluated with the DALI questionnaire: visual, auditory and tactile demand of driving activity with or without additional tasks.
- **Mental demand** is evaluated with the three tools but it is defined by different factors and is related to different activity: PSA-TLX and BMDMW evaluate driving effort (mental and/or attentional), while DALI estimates effort of attention, interference and temporal demand of either driving activity or the global situation of driving with additional secondary task (IVIS).
- **Driver's state** is assessed with the three tools in terms of psychological/mental, physical and physiological feelings.
- **Driving workload** is evaluated with the PSA-TLX and BMDMW questionnaire: they propose to estimate the performance of driving through safety consequences (risky behaviour, modification of usual behaviour, errors of driving…).

Overview of achieved experiments:

Three experiments have been conducted (see Table 1-1).

The three tools have not been tested in the three experiments as it was planned at the beginning of the task, because:

- all tools were not available at the beginning of tests and tests could not be postponed considering the time plan,
- the administration of three different questionnaires may induce potential interference between concepts evaluated and the filling procedure asked to participants (each questionnaire presents specific principles),
- the administration of the three tools is time consuming and experimental design could not afford the administration of all three tools in each experiment.

| Experimental design | CRF | INRETS | PSA |
|---|---|---|---|
| Test Environment | VR simulator | Road | Road |
| Drivers NB | 16 | 2x10 | 12 |
| IVIS | Radio | Route guidance system | / |
| ADAS | Frontal collision warning system | / | Adaptive Cruise Control (information / regulation) |
| Other tasks | | Paper map use Various messages | |
| DALI | x | x | x |
| PSA-TLX | x | x | x |
| BMDMW | x | x | / |

**Table 1-1: Experimental design of all experiments**

Considering the experimental design of the three tests, sensitivity of the three tools is evaluated through the effect of:

- the test environment : simulator vs. road
- system use : between two ADAS, IVIS vs. ADAS, IVIS + ADAS vs. IVIS or ADAS (see Table 1-2)

| Comparisons | FCWS | FCWS + Radio | ACC | Route guidance | Paper map | Various tasks |
|---|---|---|---|---|---|---|
| FCWS | | | | | | |
| FCWS + Radio | X | | | | | |
| ACC | X | x | | | | |
| Route guidance | X | x | x | | | |
| Paper map | X | x | x | x | | |
| Various tasks | X | x | x | x | x | |

**Table 1-2: Comparisons of conditions**

- for the FCWS : effect of drivers' characteristics : expert versus non-expert drivers – training/experience of simulator driving

First of all, the state of the art carried out by KITE is presented, and then main results of each experiment are presented.
The last part of the deliverable deals with the synthesis of main results gathered in all three experiments, with the main conclusion on each method and the recommendations transferred to T227.

# 2 Common plan/methodology/theoretical framework : state of the art (KITE, CRF, PSA, INRETS)

Workload is still today a loose concept, due to the many existing ways of defining it and due to the large number of different assessment methodologies that often do not consent an evident and direct comparison of recorded and collected data.

Subjective measurement has been used and accepted as the most essential method among all workload measures. This is because subjective measurement directly measures how much the subject feels in terms of workload, stress, or how demanding a task is.

When analysing the advantages of subjective methods, we have to underline that they are easy to administer and widely accepted. Subjective workload measurement doesn't require excessive instrumentation, nor does it interfere with primary tasks (depending on the moment and the frequency with which we measure it). Although there are important strengths to subjective workload measurement, we have to remember that there are also some important disadvantages. It is in fact possible that subject ratings may not reflect the real workload level or that the ratings become influenced by other biases such as dislike or unfamiliarity with the task as well as the subject's reluctance to report that things are difficult.

Even though much effort has been made to develop objective measures of workload, subjective workload assessment techniques continue to be popular due to their ease of use, general non-intrusiveness, low cost, high face validity and known sensitivity to workload variations. Subjective mental workload can be defined as a subject's direct estimate or comparative judgement of the mental or cognitive workload experienced at a given moment.

The most common method used for subjective workload measurement is requesting a subject to rate different aspects of efforts using a predefined rating scale while guided by a verbal description for each level of the scale. For this purpose, several types of uni- and multi-dimensional scales have been created.

This chapter represents a review of some of the existing scales of workload subjective assessment taken from general literature and specific domain applications (especially aviation domain). Every method is here presented in an introductive manner that can be analysed in depth by using references.

The chapter ends with a section that contains a preliminary description of the tools utilized in experimental trials of the whole work. These three tools come from automotive domain studies of workload, sometimes transferring and adapting some already existing general methods of assessment like for PSA-TLX, sometimes adapting some specific "aeronautics" workload features to drivers experienced workload as for DALI by INRETS.

There are not so many publications regarding the aforementioned methods, especially for PSA-TLX and for BMDMW. The latter is a questionnaire that found its final line-up in prevision of the experiments carried out in this phase of the project. Only DALI can account some previous validation studies that were already described in other review works of the project (see section 3).

It has to be remembered that experiments objectives were to have a feedback from subjects (drivers) about "workability" of this tools during workload assessment.

## *2.1 "OFF-LINE" ASSESSMENT METHODS*

### 2.1.1 NASA Task Load Index (TLX)

The NASA Task Load Index (Hart & Staveland, 1988) is a multi-dimensional rating procedure that provides an overall workload score based on a weighted average of ratings on six subscales: Mental Demands, Physical Demands, Temporal Demands, Own Performance, Effort, and Frustration. The table above provides a definition of each subscale:

| RATING SCALE DEFINITIONS | | |
|---|---|---|
| Title | Endpoints | Descriptions |
| MENTAL DEMAND | Low/High | How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving? |
| PHYSICAL DEMAND | Low/High | How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious? |
| TEMPORAL DEMAND | Low/High | How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic? |
| EFFORT | Low/High | How hard did you have to work (mentally and physically) to accomplish your level of performance? |
| PERFORMANCE | Good/Poor | How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals? |
| FRUSTRATION LEVEL | Low/High | How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task? |

**Table 2-1: Definitions of the six workload subscales of NASA TLX.**

The degree to which each of the six factors contribute to the workload of the specific task to be evaluated, from the raters' perspective, is determined by their responses to pair-wise comparisons among the six factors. Magnitude ratings on each subscale are obtained after each performance of a task or task segment. Ratings of factors deemed most important in creating the workload of a task are given more weight in computing the overall workload score, thereby enhancing the sensitivity of the scale.

The weights and ratings may or may not covary (for example a task could result as very stressful regarding physical demands even if physical requests might be low). Since subjects can give ratings quickly, it may be possible to obtain them in operational settings.

The Task Load Index has been tested in a variety of experimental tasks that range from simulated flight to supervisory control simulations and laboratory tasks (e.g., choice reaction time, mental arithmetic, mental rotation, etc.). The derived workload scores have been found

to have substantially less between-rater variability than one-dimensional workload ratings, and the subscales provide diagnostic information about the sources of load.
The NASA Task Load Index is a two-part evaluation procedure consisting of both weights and ratings (collected using 3 three separate computer programs that also combine the former with the latter).

The weights provide diagnostic information about the nature of the workload imposed by the task. There are 15 possible pair-wise comparisons of the six scales. The WEIGHTS software presents each pair to the subject and this one selects the member of each pair that contributed more to the workload of the task. The computer tallies the number of times that each factor was selected. The tallies can range from 0 (not relevant) to 5 (more important than any other factor).

The second requirement is to obtain numerical ratings for each scale that reflect the magnitude of that factor in a given task. Subjects respond to the six scales (presented by the RATINGS program) by marking each scale at the desired location, using either the keyboard or a mouse. Each scale is presented as a line divided into 20 equal intervals anchored by bipolar descriptors (e.g., High/Low). Ratings may be obtained either during a task, after task segments, or following an entire task. In operational situations, ratings sheets or verbal responses are more practical, while the computerized version is more efficient for most simulation and laboratory settings.

The third software program (named COMBINE program) computes the overall workload score for each subject by multiplying each rating by the weight given to that factor by that subject. The sum of the weighted ratings for each task is divided by 15 (the sum of the weights): the results are weighted workload scores, that can be graphically represented as in Fig. 2-1:

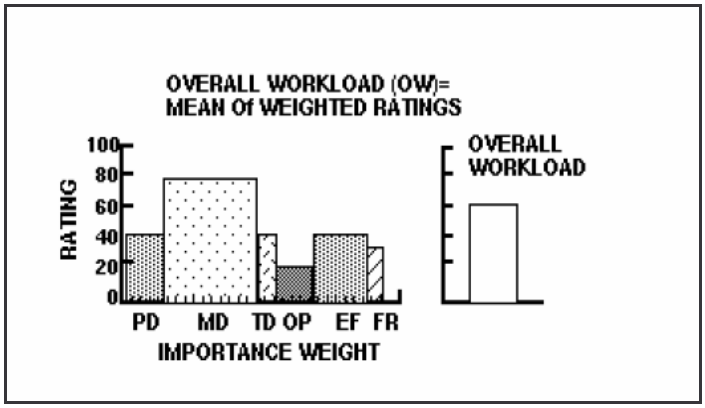



**Fig. 2-1: Graphic example of the composition of a weighted workload score.**

The bar graph on the left represents six subscale ratings. The width of the subscale bars reflects the importance of each factor (its weight) and the height represents the magnitude of each factor (its rating) in a particular task. The weighted workload score (the bar on the right) represents the average *area* of the subscale bars.
The NASA-TLX is a well known tool and it has been widely used, both in simulation as well as in field tests, in different domains. But an adaptation work seems to be necessary to exploit it in the automotive domain: some of the construct may not be so useful (as for example

"physical demand" that seem to be reduced in today vehicles), while some others could be missing.

### 2.1.2 SUBJECTIVE WORKLOAD DOMINANCE (SWORD)

SWORD (Vidulich et al. 1991) measures workload of different tasks as a series of relative subjective judgements compared to each other.
The implementation of this assessment tool is the succession of three steps:
Collect subjective between-tasks comparative ratings using a structured evaluation form (one example is given in Fig. 2-2) after the subject has finished all the tasks.

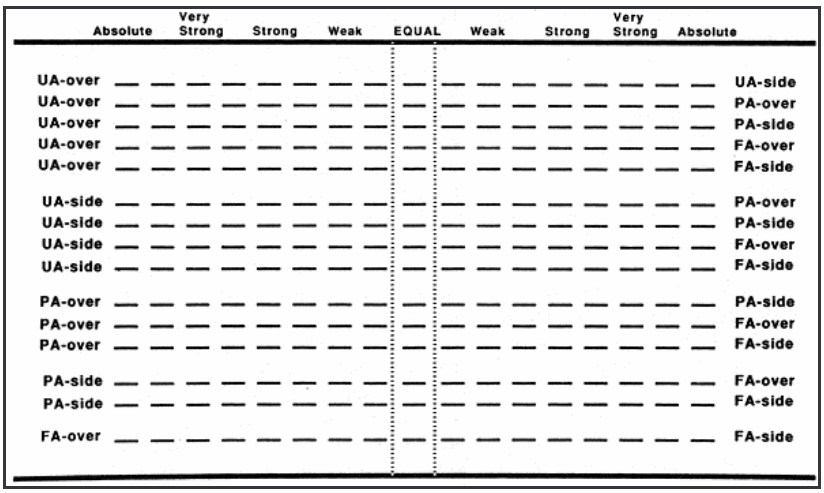



**Fig. 2-2 : An example of a SWORD evaluation form: PA, UA, etc…are the particular tasks compared each other. (reproduced from Vidulich et al, 1991).**

In a SWORD subjective evaluation form, each of all possible paired combinations of the tasks is listed in one row, with one task on the leftmost side and the other task on the rightmost side. There are 17 comparative levels between any two tasks as shown in Fig. 2-2. Namely, nine levels are assigned to each task with "EQUAL" belonging to both tasks in one row. Nine levels is regarded as a reasonable upper limit for the response scale since "seven plus or minus two items represents a pervasive limit in cognitive processing" (Vidulich et al., 1991).

If the subject perceived that the two tasks induced the same level of workload, then he marks on the "EQUAL" slot in the corresponding row of the evaluation form. If he perceived either task had workload dominance over the other, he marks a slot closer to the dominant task according to the level of workload difference.

1. Construct a judgment matrix based on the subjective ratings.

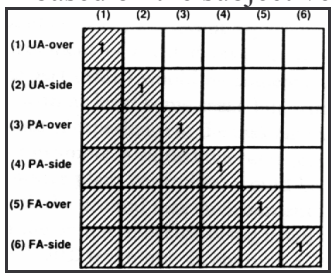



**Fig. 2-3 : an example of a SWORD judgment matrix (reproduced from Vidulich et al., 1991)**

The diagonal of the SWORD judgment matrix, as shown in Fig. 2-3, is filled with ones representing tasks that are compared to themselves. The upper right triangular area is filled with the subjective ratings (possible value are 2 to 9) of the dominance of the task in a row over the task in a column extracted from the evaluation form. The lower left triangular area is filled with the reciprocals of the numbers in the diagonally symmetric cells of the upper right area.

2. Calculate the relative ratings for each task.

The judgment matrix is subject to a consistency test. Consistency means, for example, if the rating of task A is twice as much as the rating of task B, and the rating of task B is three times as much as the rating of task C, then the rating of A should be six times as much as the rating of task C. If consistency (calculated with a specific method and indicated as $S^2$) of a subject's judgment matrix is greater than the critical value (consulted in the application method), then the judgment matrix is too inconsistent to be included in the final workload measurement.

The final rating for each task is the normalized geometric mean for that row of the judgment matrix. The rating represents a ratio scale of a task's workload level compared to all other tasks.

SWORD could be surely a good model to be exploited in automotive domain. After having divided a driving session into a sequence of tasks we could collect information (overall workload ratings) about all of these tasks, both after road tests and after simulator sessions. In this method, it is not necessary to previously define different aspects or contributing factors of workload: it is sufficient to define tasks. This could be a facilitating aspect if we would like to apply this tool to automotive domain.

### 2.1.3 SUBJECTIVE WORKLOAD ASSESSMENT TECHNIQUE (SWAT)

SWAT is a multidimensional scaling workload measurement technique. It is based on the assumption that the subject's workload can be adequately represented by three dimensions: time load, mental effort load, and psychological stress load. Each dimension has three levels: 1 signifying low load, 2 signifying medium load, and 3 signifying high load as shown in Table 2-2, which generates 27 combinations of dimensions and workload levels.

SWAT was originally designed to assess the workload associated with the operators' activities, in particular, aircraft cockpit and other crew-station environments

| *Time Load* | *Mental Effort Load* | *Stress Load* |
|---|---|---|
| 1. Often have spare time. Interruptions or overlap among activities occur infrequently or not at all. | 1. Very little conscious mental effort or concentration required. Activity is almost automatic, requiring little or no attention. | 1. Little confusion, risk, frustration, or anxiety exists and can be easily accomodated. |
| 2. Occasionally have spare time. Interruptions or overlap among activities occur frequently. | 2. Moderate conscious mental effort or concentration required. Complexity of activity is moderately high due to uncertainty, unpredictability, or unfamiliarty. Considerable attention required. | 2. Moderate stress due to confusion, frustration, or anxiety noticeably adds to the workload. Significant compensation is required to maintain adequate performance. |
| 3. Almost never have spare time. Interruptions or overlap among activities are very frequent, or occur all the time. | 3. Extensive mental effort and concentration neccessary. Very complex activity requiring total attention. | 3. High to very intense stress due to confusion, frustration or anxiety. High to extreme determination and self-control required. |

**Table 2-2 : The dimensions and scales of SWAT (reproduced from Wickens et al., 2000)**

SWAT is implemented in two phases:

- Phase 1: a scale development phase entails subjects to rate the workload of the 27 combinations of each task. A table is generated using conjoint measurement methodology to assign a workload score on a scale of 0 to 100 to each of the 27 combinations. SWAT has the capability to account for individual differences by subgrouping the subjects according to the dimensions they emphasize most in the rating. A separate workload scale can be derived for each subgroup.

- Phase 2: in an event scoring phase, subjects rate the three dimensions of a task using a scale of 1 to 3. These ratings are then translated into a workload score that was derived from the above phase.

SWAT is unique among other workload rating scales because it demonstrates interval-scale properties. Although results are similar to those obtained using the NASA-TLX, they are slightly less consistent across individuals. SWAT has been shown to be sensitive within multi-task environments and single-task laboratory settings.

Although the subjective workload assessment technique has been widely used, it has two main problems: it is not very sensitive for low mental workloads and it requires a time-consuming card sorting pretask procedure. Some variants of this methodology have been tested and matched with the original one. This has reached the objective of reducing the initial pretask operation times and preserving (or better improving) its sensitivity for low workload values. More precisely, these variations of the original instrument turned the initial card sorting procedure (CS) into a pair-wise comparison task (PWC). In this way it is possible to

- Save time at the level of phase 1
- Improve the sensitivity in phase 2 thanks to the use of a continuous scale instead of a discrete one (see Luximon and al., 2001).

SWAT was firstly designed for aircraft cockpit operations and similar tasks. By the definition of workload using 3 dimensions (everyone with 3 levels) SWAT originates 27 combinations which have to be ranked by the subject. But observing the definitions given for every dimension, they don't seem so much applicable to a driving situation, as for example "Time Load" (we should clearly define what are "interruptions" or "overlapping activities" when driving).
The definitions of "Mental efforts" seems to be inspired by the "skill-rule-knowledge" structure by Rasmussen (1987): this could be very helpful in defining driving tasks and operation sequence.
For adapting SWAT to driving situations, a better definition of workload dimensions should be carried out and given to subjects.

### 2.1.4 COOPER-HARPER SCALE and MODIFIED COOPER-HARPER SCALE (MCH).

The Cooper-Harper Rating Scale (Harper & Cooper, 1986) is a decision tree that uses

1. adequacy for the task;
2. aircraft characteristics;
3. demands on the pilot,

to rate handling qualities of an aircraft. The scale provides ordinal data that must be analyzed accordingly. The Cooper-Harper scale should be used for workload assessment only if

handling difficulty is the major determinant of workload. The task must be fully defined for a common reference.
Ratings vary from 1 (excellent, highly desirable) to 10 (major deficiencies). Non-integer ratings are not allowed (see Fig. 2-4).

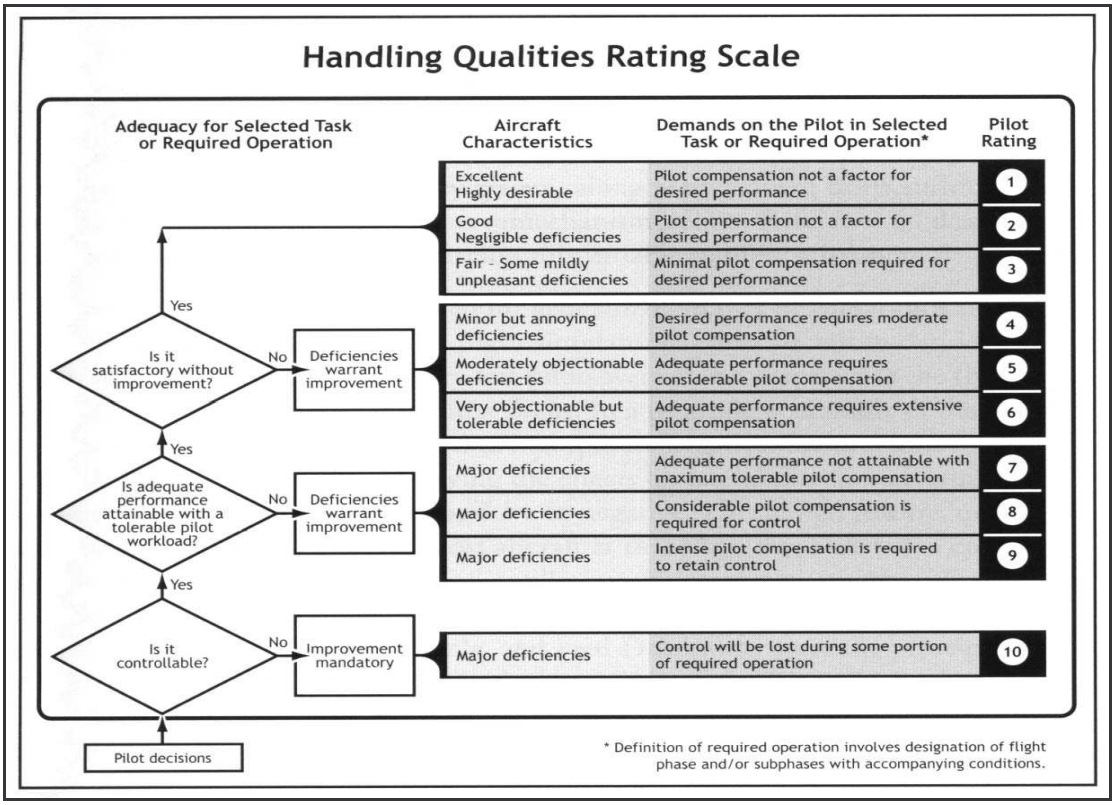



**Fig. 2-4 : The decision tree of the Cooper-Harper Scale**

Wierwille and Casali (Wierwille and Casali, 1983) noted that the Cooper-Harper scale represented a combined handling qualities/workload rating scale. They found that it was sensitive to psychomotor demands on an operator, especially for aircraft handling qualities. They wanted to develop an equally useful scale for the estimation of workload associated with cognitive functions, such as "perception, monitoring, evaluation, communications, and problem solving." The Cooper-Harper scale terminology was not suited to this purpose.

A modified Cooper-Harper rating scale (see Fig. 2-5) was developed to "increase the range of applicability to situations commonly found in modern systems." Modifications included:

1. changing the rating scale end points to “very easy” and “impossible”,
2. asking the pilot to rate mental workload level rather than controllability,
3. emphasizing difficulty rather than deficiencies.

In addition, Wierwille and Casali defined mental effort as "minimal" in rating 1, while mental effort is not defined as minimal until rating 3 in the original Cooper-Harper scale. Further, adequate performance begins at rating 3 in the modified Cooper-Harper but at rating 5 in the original scale.
Investigations were conducted to assess the modified Cooper-Harper scale. They focused on perception (e.g., aircraft engine instruments out of limits during simulated flight), cognition (e.g., arithmetic problem solving during simulated flight), and communications (e.g.,

detection of, comprehension of, and response to own aircraft call sign during simulated flight).
The modified Cooper-Harper is sensitive to various types of workloads: several studies have shown improved ratings as communication load, navigation load or the number of danger conditions increased.

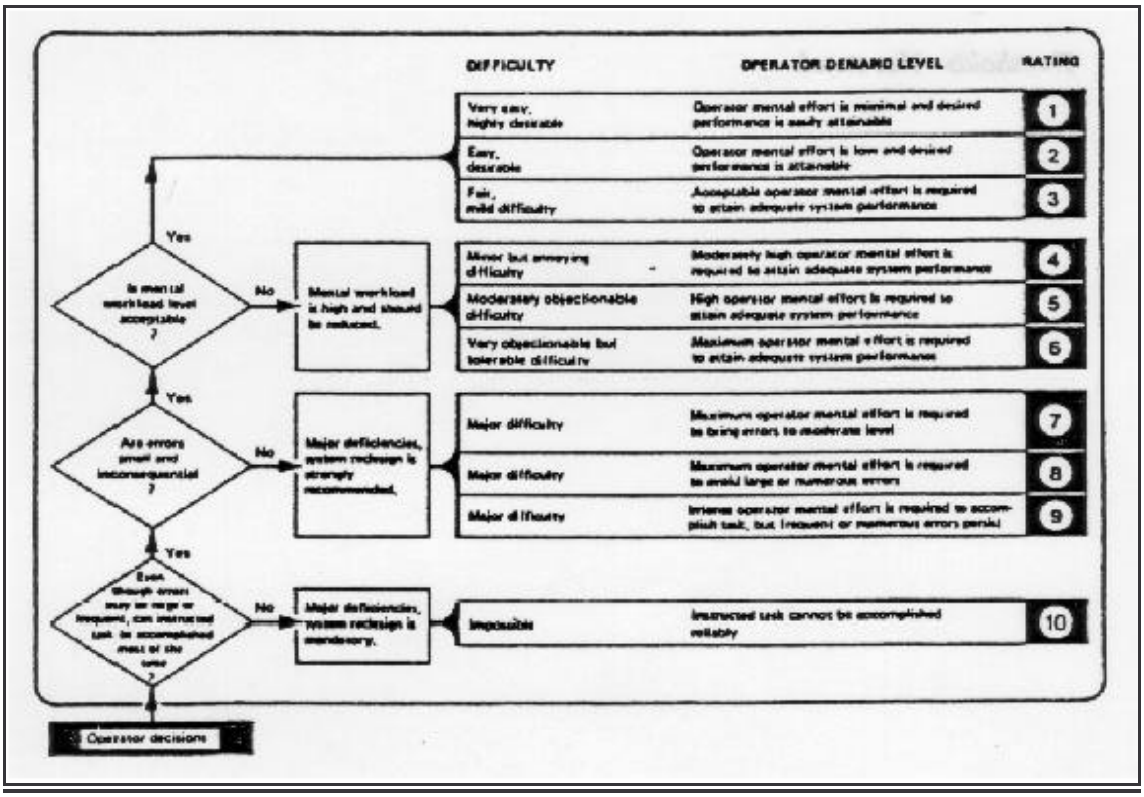



**Fig. 2-5 : The Modified Cooper-Harper Rating Scale (Wierwille and Casali 1983)**

These results suggested that the modified Cooper-Harper scale is a valid, statistically reliable indicator of overall mental workload. However, it carries with it the underlying assumptions that high workload is the only determinant of the need for changing the control/display configuration.

Finally, Wierwille and Casali (1983) recommend the use of the modified Cooper-Harper in experiments where overall mental workload is to be assessed. They emphasize the importance of proper instructions to the subjects. Since the scale was designed for use in experimental situations, it may not be appropriate to situations requiring an absolute diagnosis of a subsystem.

The MCH could be administered both at the end of an experiment as well as after a subtask.

As for previous mentioned assessment tools, also for MCH a work of adaptation and wording change should be conducted. It is what has been done for train drivers workload assessment by Rail Safety and Standards Board in United Kingdom. The Modified Cooper Harper scale (MCH) was developed to be more appropriate in complex and automated systems where operators are not required to actively control systems but are more often monitoring, perceiving, evaluating and problem solving (Wierwille and Cascali 1983). This also seems more relevant to train driving. Therefore the wording of the scale was replaced to represent activities relevant to such systems and to include task accomplishment, ability, errors,

difficulty, performance and mental workload. However, there were two main concerns for the direct transfer of the original MCH to the rail industry:

1. The wording used has been validated by pilots and may not be directly transferable to drivers
2. The tool only identifies high mental workload as unacceptable, with underload not being considered

Therefore further adaptation of the MCH was considered appropriate before introducing it to drivers. The new adapted tool created for train drivers has been called Acceptable Workload Evaluation (AWE).

The case of rail operators could be helpful for the development of a new tool for car drivers on the basis of the MCH.

### 2.1.5 MALVERN CAPACITY ESTIMATE (MACE)

MACE (Goillau and Kelly, 1996) is a quick simple and direct measure of maximum capacity. It is designed to provide a direct measure of air traffic controllers' subjective estimates of their own aircraft handling capacity.
MACE is applied at the end of a work sequence (e.g., simulation trial) and provides capacity estimates in aircraft per hour. Applications have typically been in simulation environments.
It seems difficult to find new good points that can inspire and justify its application with car drivers.

### 2.1.6 ANALYTICAL HIERARCHY PROCESS (AHP)

The analytical hierarchy process (AHP) uses the method of paired comparisons to measure workload (Vidulich, 1988). Specifically, subjects rate which of a pair of conditions has the higher workload. All combinations of conditions must be compared. Therefore, if there are $n$ conditions, the number of comparisons is $0.5n(n\text{-}1)$.

Four steps are required to use the AHP. First, a set of instructions must be written.
A verbal review of the instructions should be conducted after the subjects have read the instructions to ensure their understanding of the task. Second, a set of evaluation sheets must be designed to collect the subjects' data. An example is presented in Figure 8. Each sheet has the two conditions to be compared in separate columns, one on the right side of the page, the other on the left.

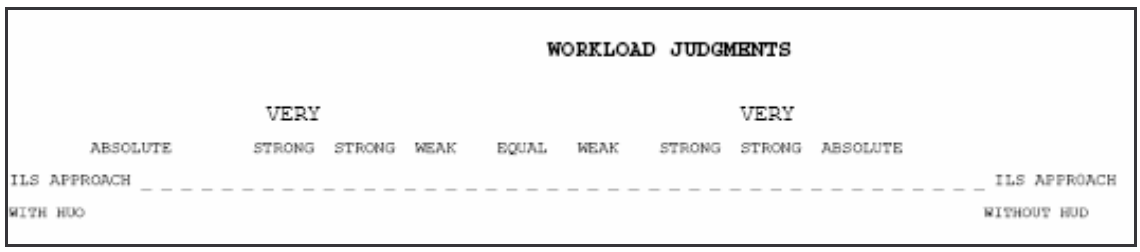



**Fig. 2-6 : AHP rating scale**

A 17-point rating scale is placed between the two sets of conditions. The scale uses five descriptors in a pre-defined order and allows a single point between each for mixed ratings (see Fig. 2-6). Vidulich (1988) defined the scale descriptors (see Table 2-3):

| | |
|---|---|
| EQUAL | The two task combinations are absolutely equal in the amount of workload generated by the simultaneous tasks. |
| WEAK | Experience and judgment slightly suggest that one of the combinations of tasks has more workload than the other. |
| STRONG | Experience and judgment strongly suggest that one of the combinations has higher workload. One task combination is strongly dominant in the amount of workload, and this dominance is |
| VERY STRONG | clearly demonstrated in practice. The evidence supporting the workload dominance of one task combination is the highest |
| ABSOLUTE | possible order of affirmation (adapted from Vidulich, 1988,p.5). |

**Table 2-3 : Definitions of AHP scale descriptors.**

Third, the data must be scored. The scores range from +8 (absolute dominance of the left-side condition over the right-side) to -8 (absolute dominance of the right-side condition over the left-side condition). Finally, the scores are input, in matrix form, into a computer program. The output of this program is a scale weight for each condition and three measures of goodness of fit.

It should be noted that the AHP is a very generic approach to measure expert judgement and expert opinion. AHP can and has been extensively utilised in many different contexts where the knowledge and opinions of experts need to be evaluated or compared in detail. Consequently, it is a very well suited to measure subjective workload and to compare among different experts.

AHP could be helpful for workload assessment in automotive domain to collect expert judgements about different situations (the conditions which have to be compared) if used in conjunction with another workload assessment tool for subjective ratings given by drivers. AHP could represent a first preliminary assessment phase of different driving situations (correlations between experts judgements and drivers ratings should be found).

### 2.1.7 CREW STATUS SURVEY

The original Crew Status Survey (Pearson and Byars, 1956) contained 20 statements describing fatigue states. The staff of the Air Force School of Aerospace Medicine Crew Performance Branch updated the original survey. It selected the statements anchoring the points on the fatigue scale of the survey through iterative presentations of drafts of the survey to aircrew members. The structure of the fatigue scale was somewhat cumbersome, since the dimensions of workload, temporal demand, system demand, system management, danger, and acceptability were combined on one scale. However, the fatigue scale was simple enough to be well received by operational crews. The fatigue scale of the survey was shortened to seven statements and subsequently tested for sensitivity to fatigue as well as for test/retest reliability. Finally, a seven-point workload scale was added.

The current Crew Status Survey (Table 2-4) provides measures of both self-reported fatigue and workload as well as space for general comments. The workload scale has been modified to enhance reliability.
The scale descriptors are:

1. Nothing to do; no system demands.
2. Light activity; minimum demands.
3. Moderate activity; easily managed; considerable spare time.
4. Busy, challenging but manageable; adequate time available.
5. Very busy, demanding to manage; barely enough time.
6. Extremely busy; very difficult; nonessential tasks postponed.

Overloaded; system unmanageable; important tasks undone; unsafe.

| NAME | DATE AND TIME |
|---|---|
| SUBJECTIVE FATIQUE<br>*(Circle the number of the statement which describes how you feel RIGHT NOW.)* | |
| 1 | Fully Alert Wide Awake: Extremely Peppy |
| 2 | Verv Lively: Responsive, But Not At Peak |
| 3 | Okay; Somewhat Fresh |
| 4 | A Little Tired; Less Than Fresh |
| 5 | Moderateiy Tired: Let Down |
| 6 | Extremely Tired; Very Difficult to Concentrate |
| 7 | Completely Exhausted: Unable to Function Effectively, Ready to Drop |
| COMMENTS | |
| WORKLOAD ESTIMATE<br>*(Circle the number of the statement which best describes the MAXIMUM workload you experienced during the past work period. Put an X over the number of the statement which best describes the AVERAGE workload you experienced during the past work period.)* | |
| 1 | Nothing to do; No System |
| 2 | Demands Little to do; Minimum System Demands |
| 3 | Active Involvement Required, But Easy to Keep Up |
| 4 | Challenging, But Manageable |
| 5 | Extremely Busy Barely Able to Keep Up |
| 6 | Too Much to do; Overloaded: Postponing Some Tasks |
| 7 | Unmanageable; Potentially Dangerous: Unacceptable |
| COMMENTS | |

**Table 2-4 : Crew Status Survey**.

These scales have been found to be sensitive to changes in task demand and fatigue but are independent of each other. Several studies has shown how ratings were different in different flight phases, demonstrating low correlations between the two estimation methods.

The aforementioned assessment tool has a great strength point: it puts into correlations 2 different constructs which have shown themselves independent in the same flight (attending to ratings) but which influence one another also in driving situations: workload and fatigue (the latter referring to the driver state more than to the physical requests of a task).

### 2.1.8 FLIGHT WORKLOAD QUESTIONNAIRE

The Flight Workload Questionnaire is a four-item, behaviorally anchored rating scale (Gawron, 2000). The items and the end points of the rating scales are: workload category (low to very high), fraction of time busy (seldom have much to do to fully occupied at all times),

how hard had to think (minimal thinking to a great deal of thinking), and how felt (relaxing to very stressful).

The questionnaire is sensitive to differences in experience and ability. For example, it has been found significant differences in the flight workload ratings between experienced and novice pilots. Specifically, experienced pilots rated their workload during an air transport flight lower than novice pilots did. However, great redundancy in the value of the ratings given for the four questionnaire items has also been found. This suggests that the questionnaire may evoke a response bias. The questionnaire provides a measure of overall workload but cannot differentiate between flight segments and/or events.

### 2.1.9 HONEYWELL COOPER-HARPER RATING SCALE

This rating scale uses a decision-tree structure for assessing, overall task workload. The Honeywell Cooper-Harper Rating Scale was developed by Wolf (1978) to assess overall task workload. It has been used successively by other researchers to assess workload associated with various Vertical Take-Off and Landings (VTOL) aircraft displays. For the small subset of conditions analyzed, the scale ratings correlated well with performance.

Subjects must answer three questions related to task performance. The ratings are ordinal and must be treated as such in subsequent analyses. The minimum point is 1, the maximum one 9 (see Fig. 2-7):

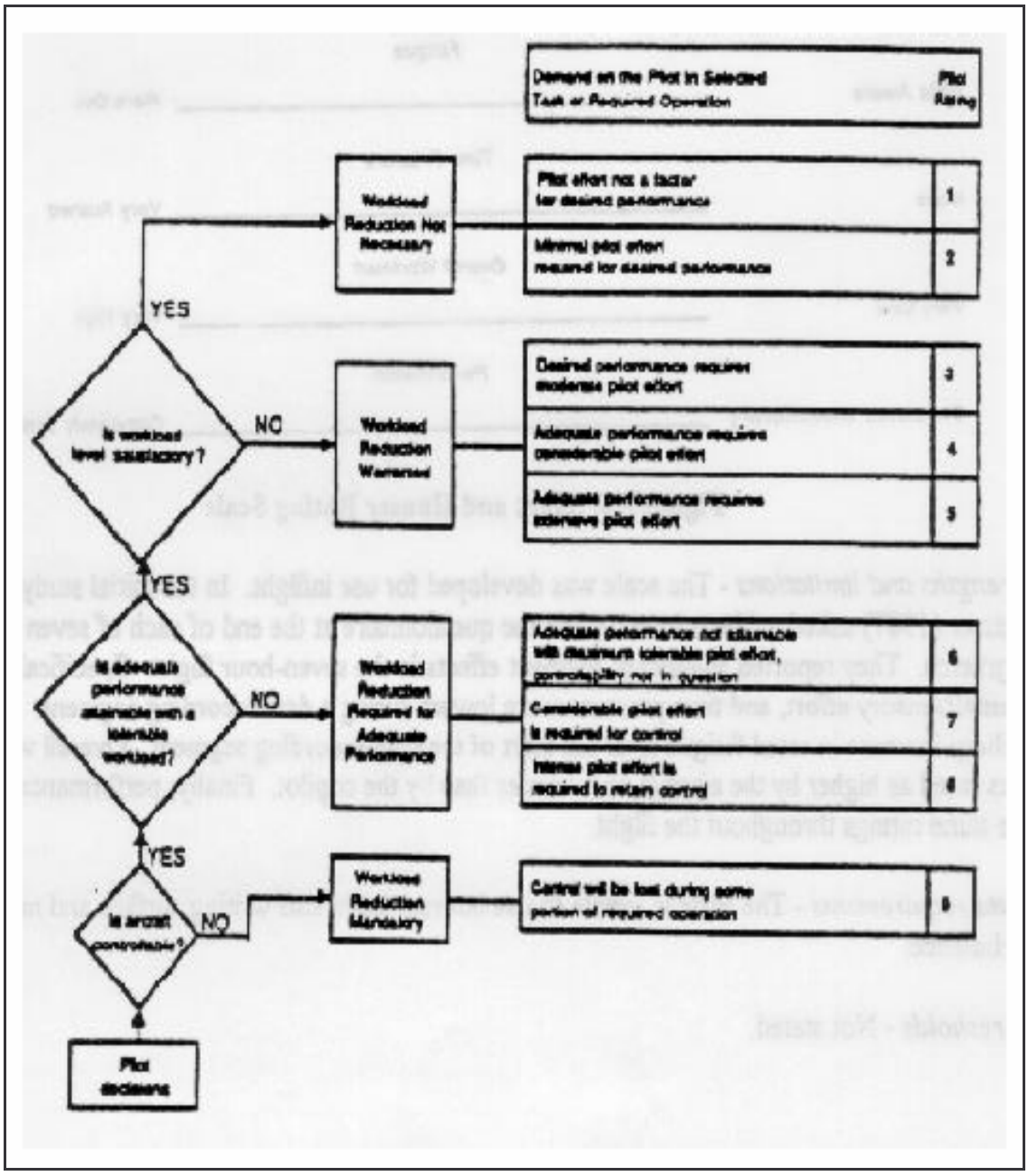



**Fig. 2-7 : Honeywell Cooper-Harper Rating Scale**

### 2.1.10 NASA BIPOLAR RATING SCALE

The NASA Bipolar Rating Scale has ten subscales (Gawron 2000). The titles, endpoints, and descriptions of each scale are presented in Table 2-5 and the scale itself, in Table 2-6 . If a scale is not relevant to a task, it is given a weight of zero.

| TITLE | ENDPOINTS | DESCRIPTIONS |
|---|---|---|
| Overall Workload | low/high | The total workload associated with the task considering all sources and components. |
| Task Difficulty | low/high | Whether the task was easy,demanding,simple or complex, exacting or forgiving. |
| Time Pressure | low/high | The amount of pressure you felt due to the rate at which the task elements occurred. Was the task slow and leisurely or rapid and frantic |
| Performance | good/poor | How successful you think you were in doing wheat we asked you to do and how satisfied you were with what you accomplished. |
| Mental/Sensory/Effort | low/high | The amount of mental and/or perceptual activity that was required (e.g., thinking, deciding,calculating, remembering, looking,searching, etc.). |
| Physical Effort | low/high | The amount of physical activity that was required (e.g., pushing, pulling, turning, controlling, activating, etc.). |
| Frustration Level | Fulfilled, Exasperated | How insecure, discourage, irritated, annoyed versus secure, gratified, content and complacent you felt. |
| Stress Level | Relaxed,Tense | How anxious, worried, uptight, and harassed or calm, tranquil, placid, and relaxed you felt. |
| Fatigue | Exhausted, Alert | How tired, weary, worn out, and exhausted or fresh, vigorous, and energetic you felt. |
| Activity Type | Skill based, Rule based, Knowledged based | The degree to which the task required mindless reaction to well-learned routines or required the application of know rules or required problem solving and decision making. |

**Table 2-5 : NASA Bipolar Rating Scale Definitions**

The number of times a dimension is selected by a subject is used to weight each scale. These weights are then multiplied by the scale score, summed, and divided by the total weight to obtain a workload score. The minimum workload value is zero; the maximum, 100.
The scale provides a measure of overall workload but is not sensitive to short-term demands. Further, the activity-type dimension must be carefully explained to pilots before use in flight.

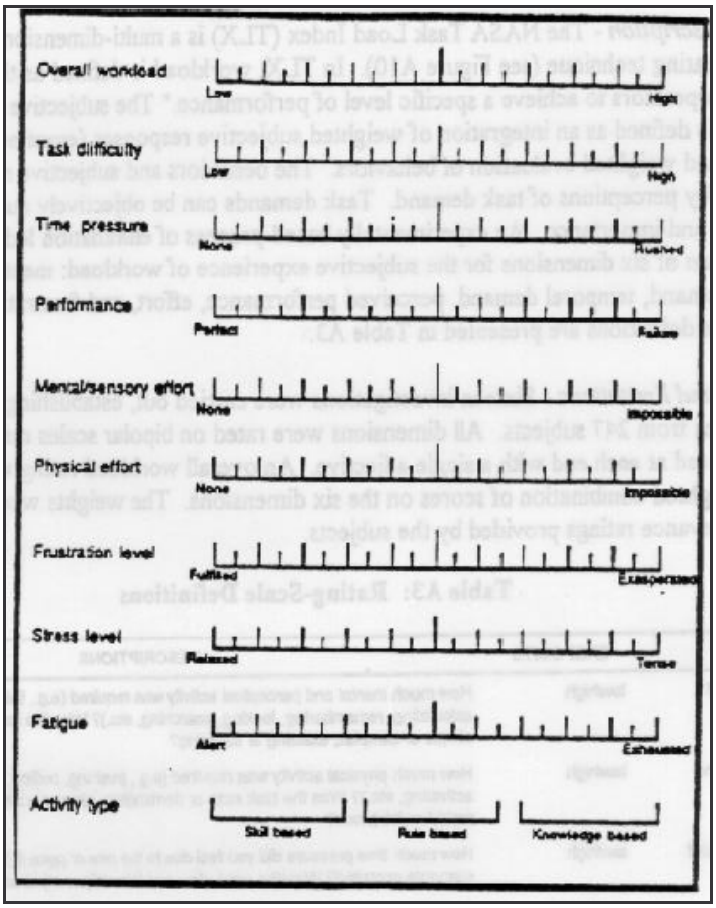



**Table 2-6: NASA Bipolar Rating Scale**

As for NASA TLX, this tool also could represent a starting point to collect overall workload ratings from drivers, after removing some variables or adding some others, more connected to the driving situation. As well as for pilots, variables should be previously explained to drivers.

### 2.1.11 EQUAL-APPEARING INTERVALS

Equal appearing intervals is a self-report technique for attitudes measurement (Gawron, 2000), very popular in psychological attitudes research, in which subjects are asked to indicate those statements in a larger list of statements (typically 20 to 22) with which they agree and disagree. Subjects' attitude scores are the average score of the scale values of the statements with which they agree.

This method can be utilised for different subjects, depending on the different purposes and objects to be assessed. Regarding workload, subjects rate the workload in one of several categories using the assumption that each category is equidistant from adjacent categories (a common assumption when using Likert-like scales). Equal intervals must be clearly defined.

### 2.1.12 MAGNITUDE ESTIMATION

Magnitude estimation is an experimental technique used to quickly and easily determine how much of a given sensation a person is having (Allard 2001).

Stevens was the first experimenter to ever suggest using magnitude estimations to quantitatively scale sensation. In a magnitude estimation experiment subjects are presented with a standard stimulus (a modulus) and are told that the stimulus has a magnitude of a certain value, like 20. The subjects are then presented with a series of stimuli that vary in intensity and are asked to assign each of the stimuli a number relative to the standard stimulus. For example, if the current stimulus is twice as intense as the standard stimulus it should be called 40 or if it is half as intense, it should be called 10.

Scaling is in no way about absolute accuracy of judgments; scaling is about the relative relationships between judgements of stimuli of different intensities. The experimenter does not use a standard stimulus in an attempt to get subjects to make judgements high in absolute accuracy…they are just trying to get subjects to use similar numbers to make results easier to interpret (Allard 2001).

Regarding workload assessment, the magnitude estimation method has been exploited: subjects are required to estimate workload numerically in relation to a standard. High correlation had been reported between workload estimates and task difficulty, has well as sensitivity comparable to estimates from the equal-appearing intervals method and SWAT. Low correlation had been found between workload estimates and reaction-time performance.

It has been suggested that the presence of a standard enhances inter-rater reliability: it is however possible that subjects may be unable to retain an accurate memory of the standard over the course of an experiment if not established and well defined.

## *2.2 "ON-LINE" ASSESSMENT TOOLS*

### 2.2.1 DEFENCE RESEARCH AGENCY WORKLOAD SCALE (DRAWS)

DRAWS (Farmer et al., 1995) is a multi-dimensional tool used to gain a subjective assessment of workload from operators. The rating scales are Input demand (demand from the acquisition of information from external sources), Central demand (demand from cognitive operations), Output demand (demand from the response required by the task), and time pressure (demand from the rate at which tasks must be performed).

DRAWS offers ease of data collection and ratings can be obtained during task performance by asking respondent to call out ratings (from 0 to 100) to verbal prompts. This can also provide a workload profile through a task sequence.

### 2.2.2 INSTANTANEOUS SELF ASSESSMENT (ISA)

Instantaneous self-assessment (ISA) is a technique that has been developed as a measure of workload (Jordan, 1992), and provides immediate subjective ratings of work demands during performance of primary work tasks such as air traffic control. Workload assessment by the individual is provided as intervals throughout the task (or in real-time simulations). The operator is prompted at regular intervals to give a rating of 1 to 5 (unidimensional scale) of how busy he/she is (1 means under-utilised, 5 means excessively busy, and relaxed, comfortable busy pace and high are the middle points respectively). These data can be used to

compare operators' perceived workload, for example, with and without a particular tool, or between different systems.
Tatterstall and Foord (1996) compared the results of ISA with those gathered from other established workload evaluation techniques. Subjective ratings of the ISA were compared to mean heart rate, heart rate variability, and error in the primary task of tracking. Results showed that the ISA was sensitive to the variations in task difficulty, as compared with levels indicated by the physiological measures, however, primary tracking performance decreased during periods where ISA responses were required. The results suggest that the effectiveness of the ISA technique is limited in comparison to less intrusive measures of mental workload (Tattersall an Foord, 1996).
Workload is here intended as amount of invested resources in a given moment. But rating could be a workload source itself, especially during excessively busy moments, when answering to a workload estimation question could be particularly difficult. It's important, for real-time evaluations, to be less intrusive as possible.

### 2.2.3 BEDFORD SCALE

The Bedford Scale (Roscoe and Allis, 1990) is a uni-dimensional rating scale designed to identify operator's spare mental capacity while completing a task. The single dimension is assessed using a hierarchical decision tree that guides the operator through a ten-point rating scale, each point of which is accompanied by a descriptor of the associated level of workload.

It is simple, quick and easy to apply in situ to assess task in high workload environments, but it does not have a diagnostic capability. Fig. 2-8 shows the procedure of application of the decision tree:

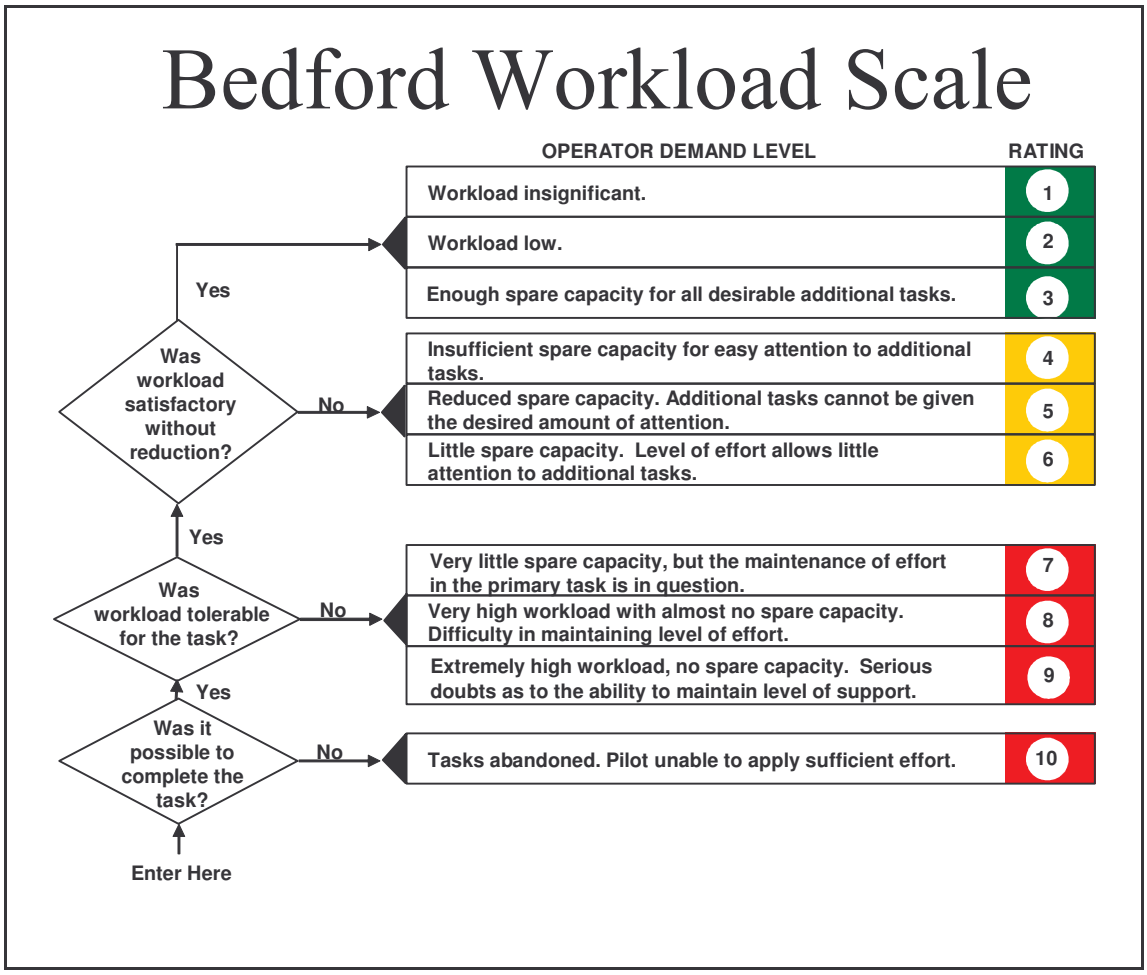



**Fig. 2-8 : The hierarchical decision tree of Bedford Scale**

### 2.2.4 DYNAMIC WORKLOAD SCALE

The Dynamic Workload Scale is a seven-point workload scale ( Fig. 2-9) developed as a tool for aircraft certification (Gawron, 2000). It has been used extensively by Airbus Industries.

| Workload | Criteria | | | |
|---|---|---|---|---|
| Assessment | Reserve Capacity | Interruptions | Effort or Stress | Appreciation |
| Light | Ample | | | Very Acceptable |
| Moderato | Adequate | Some | | Well Acceptable |
| Fair | Sufficient | Recurring | Not Undue | Acceptable |
| High | Reduced | Repetitive | Marked | High but Acceptable |
| Heavy | Little | Frequent | Significant | Just Acceptable |
| Extreme | None | Continuous | Acute | Not Acceptable Continuously |
| Supreme | Impairment | Impairment | Impairment | Not Acceptable Instantaneously |

**Fig. 2-9: Dynamic Workload Scale**

Dynamic Workload Scale is an assessment tool of workload estimation given not only by the person directly involved in the primary task operations: its ratings must be given by both a pilot and an observer-pilot. The pilot is cued to make a rating; the observer gives a rating

whenever workload changes or five minutes have passed. Two is minimum workload while eight is the maximum workload.

It has been reported high concordance between pilot and observer ratings as well as sensitivity to workload increases.

### 2.2.5 HART AND HAUSER RATING SCALE

Hart and Hauser (1987) used a six-item rating scale (Fig. 2-10) to measure workload during a nine-hour flight. The items and their scales were: stress (completely relaxed to extremely tense), mental/sensory effort (very low to very high), fatigue (wide awake to worn out), time pressure (none to very rushed), overall workload (very low to very high), and performance (completely unsatisfactory to completely satisfactory). Subjects were instructed to mark the scale position that represented their experience.

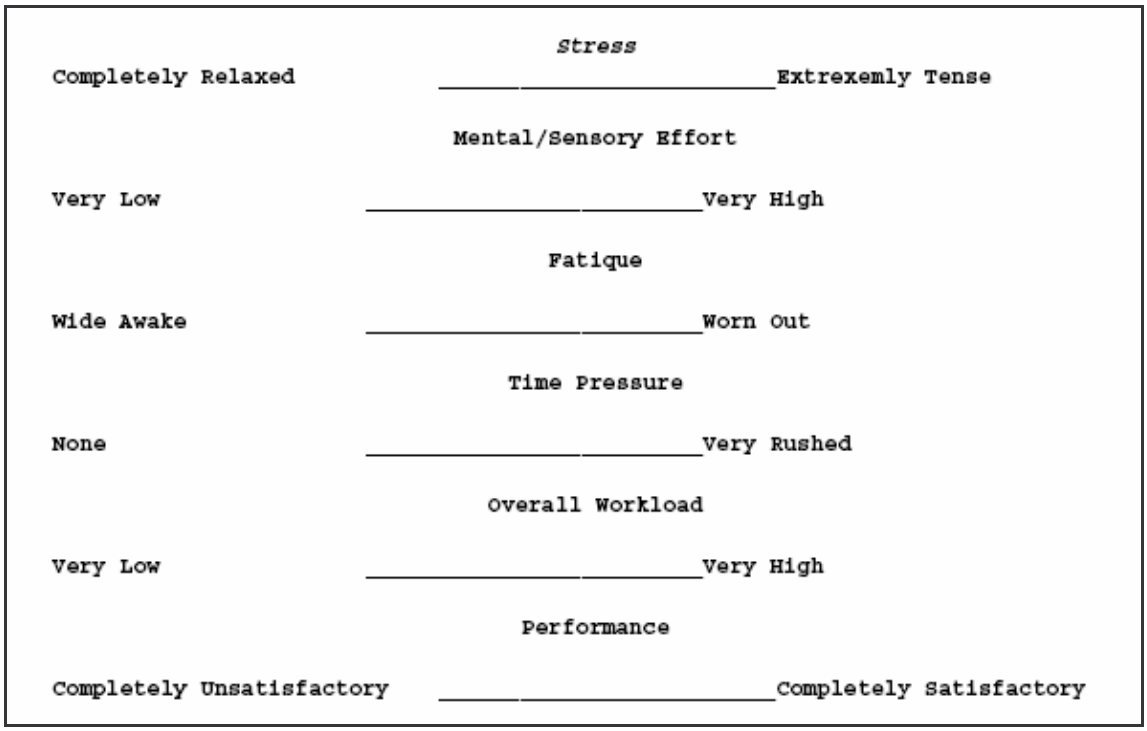



**Fig. 2-10 : Hart and Hauser rating scale**

The scale was developed for use in-flight. In the initial study, Hart and Hauser (1987) asked subjects to complete the questionnaire at the end of each of seven flight segments. They reported significant segment effects in the seven-hour flight. Specifically, stress, mental/sensory effort, and time pressure were lowest during a data recording segment. Overall workload was rated as higher by the aircraft commander than by the copilot. Finally, performance received the same ratings throughout the flight.

## *2.3 WORKLOAD PREDICTION TOOLS*

### 2.3.1 PERFORMANCE AND USABILITY MODELLING IN ATM (PUMA)

PUMA is a prototyping and evaluation tool that predicts air traffic controller workload as a function of discrete human behaviours and available human processing "resources" (Gouweleeuw et al. 1996). It is used to assess how operator workload would be affected by changes to his or her tasks.
PUMA can be applied most effectively during the concept and design phases of the life-cycle. It is a PC-based software tool that is non-intrusive to apply but has a high cost and requires training to use.

### 2.3.2 WINCREW

WinCrew is a task and workload analysis tool (Archer and Lockett, 1997). It predicts system performance as a function of human performance. It models behaviour in response to workload levels which may affect performance Fig. 2-11 shows the interface of this tool.

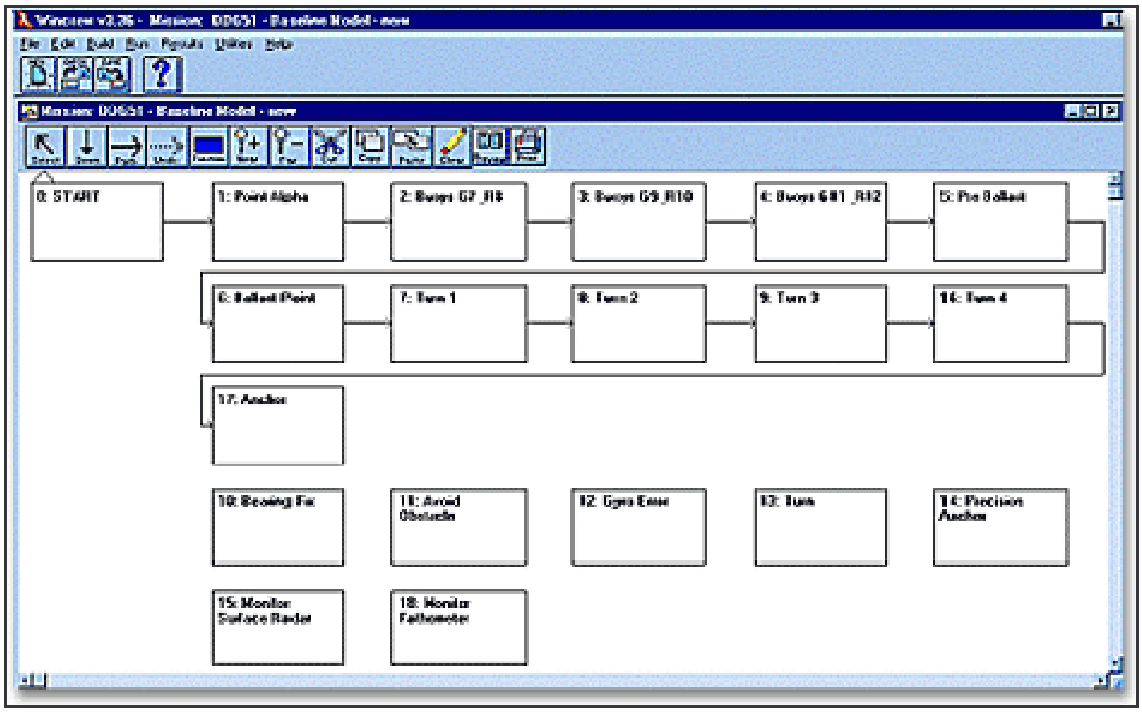



**Fig. 2-11 : The graphic interface of WinCrew.**

WinCrew predicts and assesses changes in system performance as a result of varying function allocation, number of operators or crew, level of automation, task design, mode of information presentation, and response to high workload. Through iterative use, determine high drivers affecting human and system performance. WinCrew has been used to investigate options for reduced manning, effects of different levels of automation, and workload imposed on human operators by system design concepts.

Standard task analysis data collection forms, questionnaires, etc., are used. Task times and some sequencing information come from field data, estimates from algorithms/lab studies (some available as "micromdels" in the software), and estimates from subject matter experts. Operator task assignment, task sequencing, automation concept (function allocation),

crewstation design concept (as it affects mental resources used to perform tasks), and workload associated with tasks are all derived from the system designers' concept of how the system will be used and how the human operators will perform necessary tasks.

After all data input is complete, the user selects a random number seed and executes the model in either "silent" or "animation" mode.

The following preformatted reports are produced by WinCrew as a result of model execution:

-Mission summary
-Critical path summary
-Task summary
-Operator activity
-Operator workload
-Overload
-Channel conflict
-Task timeline
-Crewstation workload
-Read user snapshot

WinCrew is a PC-software tool available on the Web with a user manual. Tool users must understand basic task analysis methods and workload concepts.

## *2.4 THE THREE TOOLS SELECTED FOR FIELD TESTS*

In this section, a preliminary description of the three methods that were chosen for workload assessment during experimental trials is presented. All these methods were previously briefly described in previous works inside the project, as it will be specified in next chapter, and they will be successively presented in and explained in details.

### 2.4.1 BMDMW

#### 2.4.1.1 Description

This section is the copy of the technical note (Di Nocera et al., 2005) describing the development of a new subjective measure specifically devoted to assess mental workload in driving tasks (see independent Appendix I.)

Mental workload has been defined as a multidimensional construct, and a great deal of studies has focused on the investigation of the general factors underlying the investment of resources in executing a task. Among the others, subjective measures are often used to gather information about operator workload. They are relatively easy to use, inexpensive, and provide fast assessments by asking the user to rate general aspects such as "temporal demand" or "performance" (as it happens using the NASA-TLX). However, two main criticisms can be addressed to this type of measure. First, they leave to the users the difficult job to relate such general aspects to the specificity of the task under evaluation. Second, they are often unrelated to behavioural and physiological indexes of mental load (Annett, J., 2002). With this in mind, albeit this approach has permitted to devise general measures of mental workload, it is questionable whether this can be considered as a proper measurement procedure. Indeed, even if there is no reason for considering the mental workload experienced by an aircraft pilot qualitatively dissimilar to that experienced by a car driver, the subjective evaluation of workload may benefit of a contextualization. Most of the tasks under evaluation have their

own specificity, as well as known variables affecting the operator load. Thus, the lack of correlation between subjective and behavioural measures often found in the literature may be due to the abstractness of the questions asked. To overcome these issues, the Behavioural Markers of Driver Mental Workload (Di Nocera et al, 2005) has been proposed as a subjective measure based on ratings of behavioural patterns related to driving tasks. This tool is a questionnaire using a 5-point Likert scale ("never" to "very often"). Currently, a set of 42 items is under evaluation for selection. The final version of the questionnaire will include 34 items. An extensive literature review has been realized to the aim of clarifying the behavioural effects of mental workload in driving tasks and their reliability. This review was the basis for item generation. Particularly, perceived and favourite speed have been reported to be related to mental load (Couyoumdjian et al, 2002) and this issue was included in several items (for example, "my speed was lower than I desired"). Furthermore, since drivers may tend to drive slowly and cautiously to compensate cognitive overload, specific items related to speed decrease were included in the questionnaire ("I have been reducing speed"). Items related to lane excess (vehicle out of lane) and lane deviation (tracking errors) were also included ("I have been adjusting my position respect to the lane") because these effects were found to be related to mental workload (Lansdown et al, 2002). Since the specific experimental needs, some BMDMW items were not administered. Those items refer to urban driving and were not applicable to the simulation study we run. The authors of the tool suggested to collectively label them as "urban environment module". The following items were translated by CRF from the original Italian version.

The BMDMW final questionnaire can be found in independent Appendix I.1: BMDMW Questionnaire.

#### 2.4.1.2 Definition of Preliminary BMDMW Factors

University of Rome administered the questionnaire to 160 participants. In particular, participants had to fill the questionnaire in by evaluating their experience of driving during the previous six months.
A Factorial Analysis study (Terenzi et al, 2005) done on this data showed the existence of a six-factor structure accounting for the 42% of common variance.

Starting from the entire questionnaire, the following items did not result to fit in any factor in Factorial Analysis. They are, therefore, rejected in the final version of the questionnaire:

- Reach destination without paying attention to path
- Look at the speedometer
- Slow down unwisely
- Use arrows to indicate direction changes
- Overtake close to a curve
- *Slow down next to a crossroad*
- Flap eyelids
- *Suddenly realize that pedestrians are nearby*

Below a brief description of Factors with a list of all items saturating that factor is provided. Items in *italic* are those belonging to the "Urban environments" module of the questionnaire.

**Factor I: Disengagement**

Items in this factor refer to disengagement behaviour during driving task

- Forget where to turn

- Slow down after watching speedometer
- Adjust vehicle trajectory according to lane
- Restart without putting the car in the first gear
- Suddenly understand to run faster than usual
- *Have no difficulty in deciding what road to take at crossroads*
- *Have no difficulty in deciding what road to take at roundabouts*

**Factor II: Vehicle Monitoring**
Items in this factor refer to Vehicle Monitoring during driving task

- Look at rear-view mirrors
- Roughly steer
- Watch surrounding landscape
- *Be able to orientate in little known roads*

**Factor III: Route Monitoring**
Items in this factor refer to Route Monitoring during driving task

- Have shrunken Neck muscles
- Look at the odometer
- Yawn
- Slow down in case of low visibility (fog, strong rain)
- Slow down after watching speedometer
- *Slow down at zebra crossings, even if no pedestrians were present*

**Factor IV: Road Awareness**
Items in this factor refer to Situation awareness loss specifically for Driving task

- Do other things during driving (tune the radio, smoke, answer the phone…)
- Watch over other vehicles
- Overtake with continuous lane
- Overtake
- Have no difficulty in deciding whether go straight or turn
- Accelerate unwisely
- Look at the odometer
- *Slow down when the Traffic Light is yellow*
- *Accelerate when the Traffic Light is yellow*

**Factor V: Control**
Items in this factor refer to Vehicle Control

- Slow down next to curves
- Keep safety distance from other vehicles
- Do other things during driving (tune the radio, smoke, answer the phone…)
- Look at spies (fuel, water, oil)
- Be overtaken
- *Accelerate at a crossroad*

**Factor VI: Fatigue**
Items in this factor refer to symptoms of physical fatigue related to Driving task

- Feel eye sting
- Adjust vehicle trajectory according to lane

- Feel eyelid heaviness
- Drive outside your lane
- Slow down when a vehicle from the opposite direction gets close
- Suddenly understand to run slower than usual

Each item was rated using a 5-point scale, from “Never” (marked as 1) to “Often” (marked as 5)

### 2.4.2 DALI

#### 2.4.2.1 Definition

The Driving Activity Load Index or DALI is a tool allowing the subjective evaluation of the mental workload (See independent appendix II.2) and specifically developed for the driving context. It is coming from an adapted version of the NASA-TLX developed by Hart & Staveland in 1988.

The NASA TLX method assumes that the workload is influenced by 6 main factors, namely men*tal demand, physical demand, temporal demand, performance, frustration level and effort*.
After assessing the magnitude of each of the six factors on a scale, the individual performs pair-wise comparisons between these six factors, in order to determine the higher source of workload factor for each pair. A composite note quantifying the level of workload is set up by using both factor rating and relative weights computed from the comparison phase.

The NASA-TLX has been set up by the NASA about 20 years ago, mainly in a context of airspace requirements. The factors identified at that time were adapted for the context. Nevertheless, some modifications are required in order to fit in a more adequate way with the driving context:

- the **Physical** component is defined as: " How much physical activity was required? - pushing, pulling, turning, controlling, activating,...-"
  This question is not very relevant when considering the driving activity where the control of the vehicle is quite automatic for an experienced driver, and where manoeuvres are not supposed to be physically demanding in our nowadays modern cars.

- the **Mental** component is defined " How much **Mental** and perceptual activity was required ? - thinking, deciding, calculating, remembering, looking, searching,...-".
  This statement covers both perceptive and cognitive aspects of the workload, and it would be interesting in the context of the driving task to be able to identify these various modalities.

- the **Performance** factor « What is the level of performance for the task »
  The level of performance can be evaluated using objective data coming from the driver behaviour and from the driving activity.

Furthermore, the subjective rating of a good performance by the driver can show discrepancies with the measured one, but this difference might be due to many other factors than the mental workload itself - low or high self-esteem, motivations to fit to the standard performance,...-

The objective of the DALI method (Pauzié, 1994) is to evaluate the workload during a well-defined task, namely the driving task. A specific focus has been devoted to the fact that this driving context might include the implementation of an in-vehicle system (factor Interference). Usually, the DALI results are compared between a situation considered as "normal driving" and a situation with a specific equipment supposed to influence the driver's workload in a positive or in a negative way.

The method has been to redefine the main factors describing the components of the driver's workload and to keep the way of computation for the score of the NASA-TLX.

The process has been to ask scientific experts and ergonomists involved in the driving task studies to define which were, in their opinion, the main factors inducing mental workload for people driving a vehicle equipped with an on-board system (car phone, driving aid system, radio...). The results of this investigation allowed to set up the following factors:

| Title | Endpoints | Description |
|---|---|---|
| Effort of attention | Low/high | To evaluate the attention required by the activity – to think about, to decide, to choose, to look for… |
| Visual demand | Low/high | To evaluate the visual demand necessary for the activity |
| Auditory demand | Low/high | To evaluate the auditory demand necessary for the activity |
| Tactile demand | Low/high | To evaluate the specific constraint due to the tactile stimulation during the driving activity |
| Temporal demand | Low/high | To evaluate the specific constraint due to timing demand during the driving activity |
| Interference | Low/high | To evaluate the possible disturbance between the driving activity and any other supplementary task such as phoning, using systems or radio. |
| Situation stress | Low/high | To evaluate the level of constraints / stress during the driving activity such as fatigue, insecure feeling, irritation, discouragement,… |

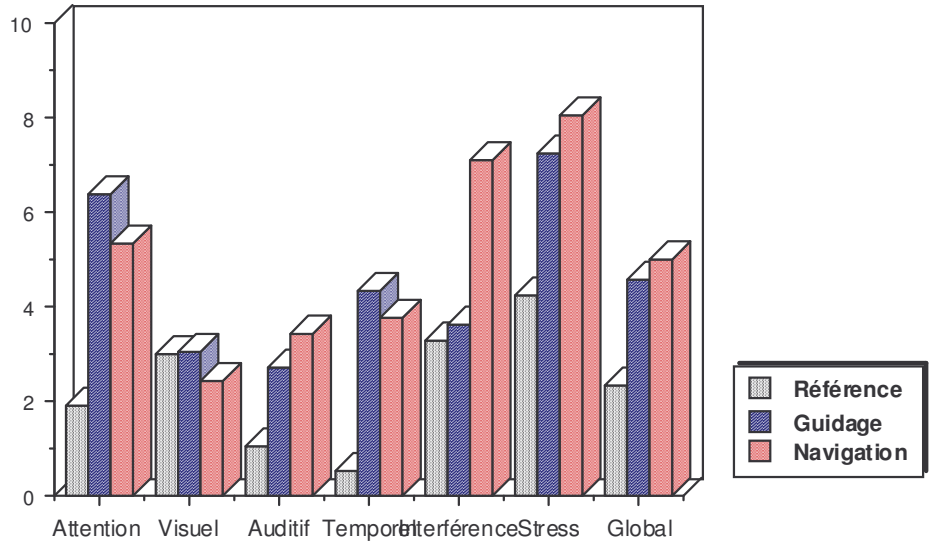
**Table 2-7 Description of DALI's factors**

After the rating procedure for each factor on a scale from 0 to 5, the paired comparison procedure is conducted in the same way as in the TLX scoring. The computation of the weighed rate for each of the six dimensions allows getting a score corresponding to the subjective evaluation of the workload factor **by factor for the 6 dimensions** and a **global** score which corresponds to the average of the 6 factors.

#### 2.4.2.2 Previous Results

Evaluation of a Guidance/Navigation System (Pauzié, 1994) see Fig. 2-12

- The gl obal effect of workload for both situations would be due to a **lack of familiarisation** concerning the use of the system.
- There is an effect of **poor temporisation** for **auditory messages**




There is an effect of **higher interference** with the driving task for **Navigation** in comparison with Guidance

**Fig. 2-12: DALI-Evaluation of a Guidance/Navigation system with the DALI questionnaire**

Results from the evaluation of a Mobile Phone System (Pauzié & Pachiaudi, 1997)

- There is an effect of **auditory demand** and **interference** with the driving task
- There is no effect of visual demand.

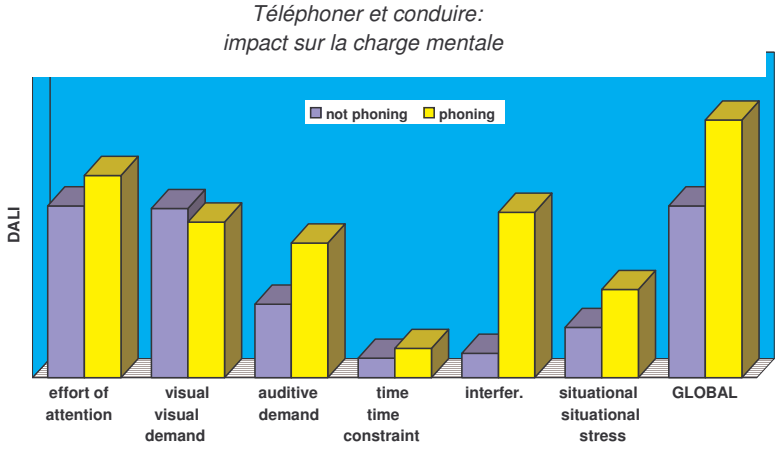



**Fig. 2-13: DALI-Evaluation of a Mobile Phone System with the DALI questionnaire**

In addition to the 6 factors used in the previous studies (**Effort of attention, Visual demand, Auditory demand, Temporal demand, Interference, Situational stress**), a supplementary factor: **Tactile demand**, has been used in the framework of the experimentation conducted in the AIDE project. Nowadays, proprioceptive perception is not very well known in the context of the driving task, and there are more and more projects about haptic systems for the driver. In this experiment, the objective was to investigate how this stimulation is perceived by the driver in comparison with the auditory and the visual ones. We made the hypothesis a priori that theoretically, tactile stimulations are not inducing high level of mental workload. In this experiment, the use of this stimulation was an opportunity to evaluate the subjective evaluation tools for this specific case.
After the rating procedure for each factor, the paired comparison procedure is led in the same way as in the TLX scoring. The computation of the weighed rate for each of the six dimensions allows getting a score corresponding to the subjective evaluation of the workload **by factor** and a **global** score.

The figure below shows that DALI's tool is dedicated to evaluate various components of driving task workload: perceptual load, mental workload and driver's state.

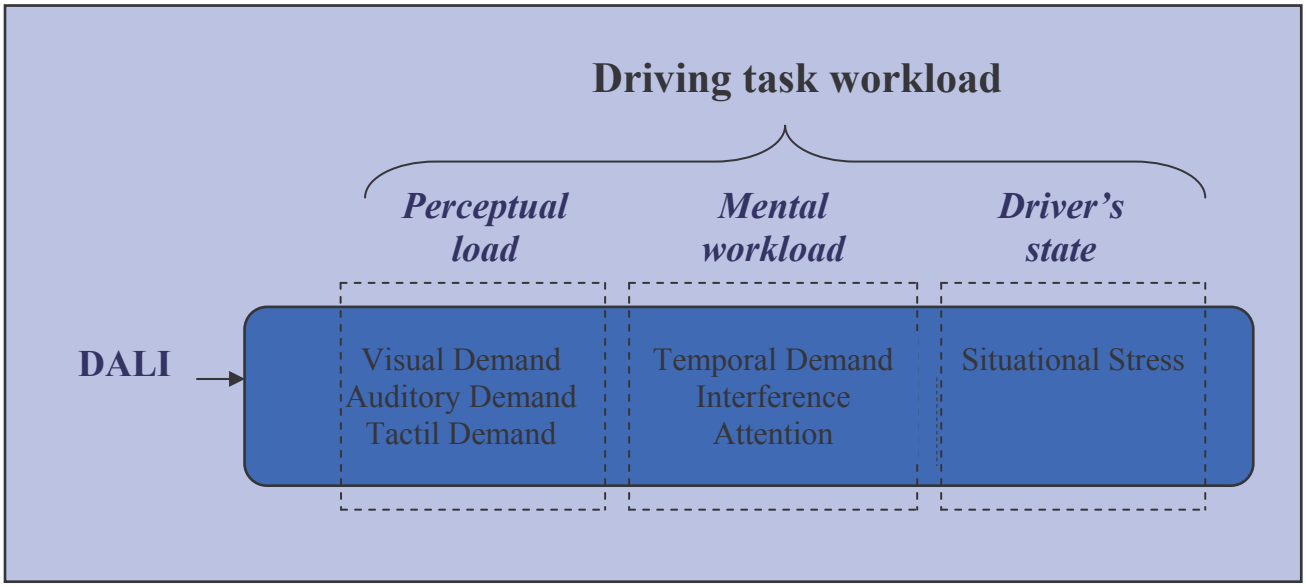



**Fig-2-14 : Overview of the DALI 's tool to evaluate the various components of the driving task workload**

#### 2.4.2.3 Administration of the DALI

The tool administration happens immediately after the conclusion of session under evaluation. The experimenter can use a paper version of the questionnaire and make it filled in by the driver (see the complete version in Independent appendix II.2). A software developed by INRETS/LESCOT is also available: the driver enters the answers directly on a lap top; the computation of the global scores and the weighted scores can be done afterward automatically without any additional task for the experimenter.

#### 2.4.2.4 Computation of the DALI

$$W_i = \alpha_i R_i$$

- Computation of the weighted DALI for each factor

Ri = li X 20, where li is the value estimated on the scale for each factor (we propose to the subject a scale with 6 points (from 0 to 5), allows to get values from 0 to 100.

Pair comparison
αi = Ci/(n-1) with Ci is the number of times a given factor has been selected during the pair comparison where factors are compared two by two, and n-1 corresponds to the number of pairs for this factor (no comparison of the factor with itself) with n= number of factors (n=7 for INRETS).

- Computation of the global weighted DALI

$$Wglobal = (\sum_{i=1}^{n} \alpha_i R_{i)/n}$$

The global weighted DALI score is the mean of the weighted score of each factor.

#### 2.4.2.5 Data analysis

Driving Activity Load Index is a method to evaluate perceptual demand and mental workload. 7 factors are evaluated, drivers are asked to choose one score between 0 to 5, corresponding to the level of constraint felt (see Table 2-8).

| DALI | Raw scores | Weighted scores |
|---|---|---|
| Effort of attention | Mean of group<br>Standard error of group<br>Correlation (Bravais-Pearson) between raw scores and weighted scores | |
| Visual demand | | |
| Auditory demand | | |
| Tactile demand | | |
| Temporal demand | | |
| Interference | | |
| Stress | | |

**Table 2-8 : DALI data processing**

### 2.4.3 PSA-TLX

PSA-TLX is a driving task oriented method, designed for real road driving (see independent appendix III.2).

The method is dedicated to estimate:

- Perceived driving workload: impact and compatibility of ADAS / IVIS use with driving safety, in terms of efficiency of the compromise between supplied mental effort and reached level of performance,
- Considering the driver state - stress, fatigue, dissatisfaction and discouragement.

PSA-TLX has been designed by the Perception and Human Factor Research department of PSA Peugeot-Citroen, following to the acknowledgment of missing sensitive subjective method to evaluate workload in automobile field (see Gautheret and Nathan 2001).
Indeed, the most frequently used method NASA-TLX or -RTLX (without factors weighting) does not present acceptable level of sensitivity and has serious difficulties of scores interpretation.
All issues known about this tool come from the context this method has been designed: this method has been initially designed in aeronautics, and consequently is not adapted to the driving task, as task's demands differ basically.
Methodological and theoretical issues emphasized in many previous studies deal with the way to compute workload index, drivers' difficulty for scoring considering the formalism of the scale, the unrealistic postulate of nil level of workload.

Major points emphasized in a state of the art achieved on workload subjective methods ((see Gautheret and Nathan 2001) have led to specify the main characteristics of a subjective tool dedicated to estimate workload for driving task's and to help in IVIS/ADAS design.

Thus, PSA-TLX has been designed to fit with these requirements:

- The method is oriented to driving, driving activity is evaluated whatever additional tasks or embedded systems use.
- Workload is a multidimensional and complex concept, as it results from compromise between allocated effort and reached performance, considering that task's demands are changing along the activity, as being influenced by reached performance and driver's state.
- Driving performance is evaluated in terms of errors defined subjectively (see section 2.1.4),

#### 2.4.3.1 Driving task model : Schlegel (1993)

Driving is evaluated at a global level and through different sub-tasks defined from Schlegel model of driving task's components

| Driving task model of Schlegel (1993) | | PSA-TLX driving sub-tasks |
|---|---|---|
| Vehicle control/guidance | Path Maintenance, Speed adaptation , Reactivity to environment (situational performance)<br>• High level perceptual processing linked to three sub-tasks achieved in parallel.<br>• High level of visual attention | Control of the lateral motion<br>Control of the longitudinal motion<br>Itinerary following |
| Navigation | Choice of itinerary : planning and search of direction<br>• Mental representation of the planned itinerary, extraction of information from a map, instructions… | Itinerary following |
| Decision making | Events detection behavioural adaptation to environment<br>Vehicle control/guidance (overtaking, speed adaptation approaching an intersection…)<br>Navigation : choice of itinerary<br>• Perception and attention, cognitive processes | Reactivity to dynamic road environment<br>Reactivity to static road environment<br>Itinerary following |
| System handling and supervision | Supervision of vehicle state through warnings and indicators detection<br>• Attention (visual, auditory, haptic…) | Use of driving controls<br>Reactivity to in-vehicle status and safety environment |
| Communication | Conversation with passengers, phone use.<br>• Perception and attention resources, cognitive processes | *Not kept in the assessment: during test, conversations are reduced to minimum with experimenter, no additional passenger, no conversation on phone.* |
| Driver's state | Fatigue, stress… | Stress<br>Fatigue<br>Overall dissatisfaction discouragement |

**Table 2-9 - Correspondence between Schlegel's model and PSA-TLX driving task components**

### 2.4.3.2 Mental workload model : De Waard (1996)

PSA-TLX purpose is to evaluate workload using the model of De Waard (1996) where workload results from a compromise more or less efficient, between effort supplied to achieve the task and performance reached in task achievement and driver's state; the efficiency depends on task difficulty (level of demand) (see Fig. 2-15)

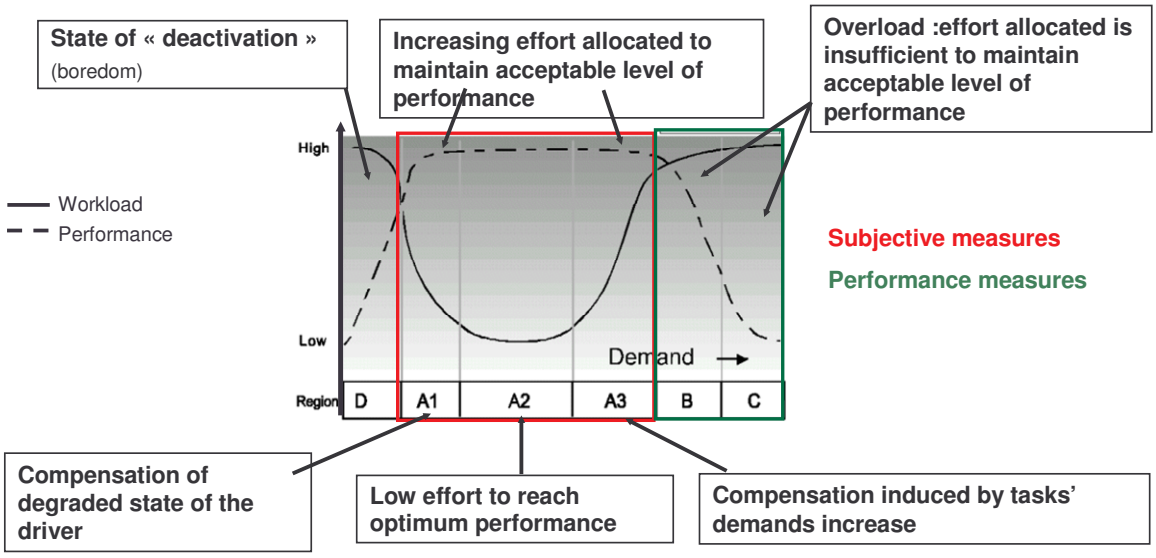



**Fig. 2-15 : De Waard Model of mental workload (1996)**

De Waard defined mental workload as resulting from the compromise between effort allocated for a task and the reached performance, with the assertion of non linear correlation between effort and performance, because the third factor "driver's state", depending and redefining task difficulty (demands)
According to tasks' demands and driver's state, mental workload induced by a task may be assimilated to the effort required and allocated to reach an acceptable level of performance which is the goal of performance a driver has for driving. High level of effort may be a compensation of increased task's demands or degraded state of the driver.
Effort supplied by the driver to achieve the task results from control between task's demands, task's performance and driver's state, knowing that these three parameters have reciprocal influences.

De Waard positions and justifies the particular relevance of subjective measures in situations with high level of demands but without any visible degradation of performance: increased effort reflects increase mental workload.

#### 2.4.3.3 Components of PSA-TLX method

PSA-TLX aims to evaluate driving workload through mental effort and performance safety and driver's state (see Table 2-10).

| Driving activity | | Driver's state |
|---|---|---|
| **Performance safety** | **Mental Effort** | |
| Overall driving | | Stress |
| Control of the lateral motion | | Fatigue |
| Control of the longitudinal motion | | Overall dissatisfaction |
| Ability to react to dynamic road environment | | Discouragement |
| Ability to react to static road environment | | |
| Itinerary following | | |
| Use of controls and driving equipments | | |
| Ability to react to safety and status indicators inside vehicle | | |

**Table 2-10 : PSA-TLX components of workload**

#### 2.4.3.4 Instruction given to drivers

The evaluation of each test-condition is performed just after driving, by comparing with one same familiar situation chosen by the driver himself where he felt himself at ease, with no stress, achieving the task with a good compromise between effort and performance.
The objective is to estimate driving workload in one test-condition in comparison with an unique reference situation i.e."low level of workload induced by a comfortable driving situation to reach acceptable performance", considered as subjective and assumed to be not nil and to vary according to individuals.
This unique reference situation can not be a no-system condition (baseline driving), because this condition of driving without any system induces workload inherent to the experimental and artificial situation.
In that way, the interest here is to evaluate driving performance compared with individual and subjective perception of risky driving (variable from one individual to another),
PSA-TLX aims to point out the "relative" measure of workload and the pertinence to keep individual specificity.

#### 2.4.3.5 Scale, meaning and scoring

Following to driver's difficulty for scoring considering the absence of meaning of scores to select, PSA-TLX proposes to use 6 levels of meaning in a scale range [0 ; 100], with a resolution of 5 points (SeeFig. 2-16) :

- to avoid any potential implicit and subjective meaning of scores,
- to have a safety threshold that can be use as a decision criteria to reject or not a system : threshold
  - for performance : 55 (any score > 55 reveals high disruption and many serious errors show unsafe performance),

- for effort : 20 (any score < 20 if driver's negligence is proved) and 55 (any score > 55 reveals high effort incompatible with safe driving as may have impact on driver's state and driving performance).

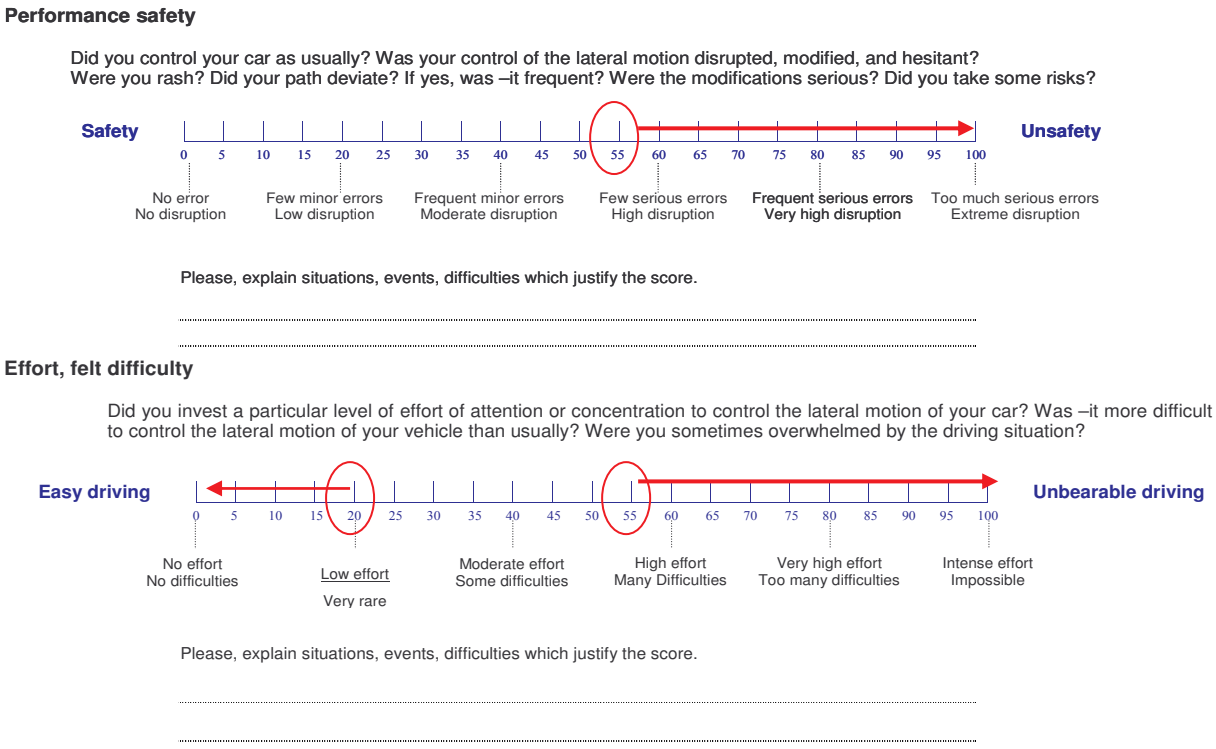



**Fig. 2-16 : PSA-TLX scale (example of lateral control evaluation)**

### 2.4.3.6 Data processing and analysis

As workload is a relative concept and as safety threshold are proposed, data processing and analysis is not based only in group tendency and statistical analysis (see Table 2-11)
Individual scores are analysed at a descriptive level with taking into account the safety threshold (see technical document in appendix III.3)

Data processing and analysis has two major parts:

- independent analysis for effort and performance
- analysis of relation between effort and performance (compromise operated in the activity).

The driver's state is currently not analysed at the same level because of the gap to integrate this third dimension from DeWaard workload model. This dimension influence task's demands and consequently workload, but is not taken into account in effort and performance of driving in data processing.

| | | |
|---|---|---|
| Driving Performance safety | Global score | Group trend :<br>Mean and standard error<br><br>Individual scores |
| | Lateral control | Group trend :<br>Mean and standard error<br>Distribution of scores<br>Performance management<br><br>Individual scores |
| | Longitudinal control | |
| | Reactivity to dynamic road environment | |
| | Reactivity to static road environment | |
| | Itinerary following | |
| | Use of driving controls | |
| | Reactivity to in-vehicle indicators | |
| Driving effort | Global score | Group trend :<br>Mean and standard error<br><br>Individual scores |
| | Lateral control | Group trend :<br>Mean and standard error<br>Distribution of scores<br>Effort management<br><br>Individual scores |
| | Longitudinal control | |
| | Reactivity to dynamic road environment | |
| | Reactivity to static road environment | |
| | Itinerary following | |
| | Use of driving controls | |
| | Reactivity to in-vehicle indicators | |
| Efficiency of the compromise effort / performance | Lateral control | For each driving sub-task, NB of drivers for each compromise :<br>- optimal<br>- very favourable<br>- favourable<br>- unfavourable<br>- very unfavourable |
| | Longitudinal control | |
| | Reactivity to dynamic road environment | |
| | Reactivity to static road environment | |
| | Itinerary following | |
| | Use of driving controls | |
| | Reactivity to in-vehicle indicators | |

**Table 2-11 - PSA-TLX data processing**

Statistical analysis is based on non parametric test (Wilcoxon and Spearman) as the total number of drivers in each group is relatively low (< 12).

#### 2.4.3.7 Results of previous studies

Before AIDE project, PSA-TLX has been tested on road tests to evaluate ADAS (night vision enhancement system) and IVIS (dialling and destination selection with voice recognition and no visual display).
Results showed that PSA-TLX is sensitive to different levels and natures of task's demands – visual/cognitive demands of a night vision enhancement use, auditory and cognitive demands of IVIS tasks (dialling and destination selection), and the seven driving sub-tasks present different levels of sensitivity.

ADAS : night vision enhancement system
Results have been presented in National Congress of French Society of Psychology 2005, special session "Driving assistance and road safety" (see Nathan F. and Chin E., 2005).

10 young male drivers (with same high driving experience and familiar with night driving) have driven without the system during 10 minutes after a familiarisation phase, and then with the night vision system during 15 min, on non lighted secondary road. The night vision system is equipped with a far infra-red camera, displaying a black and white image in a embedded screen, with visual warning when a pedestrian is detected (see Chin E. and Nathan F. 2004).

Group trend indicates good and acceptable driving performance estimated by drivers: learn to use a night vision enhancement system leads to low disruption (see Fig. 2-17). Important variability between drivers is observed. No significant difference appears between both conditions of driving.

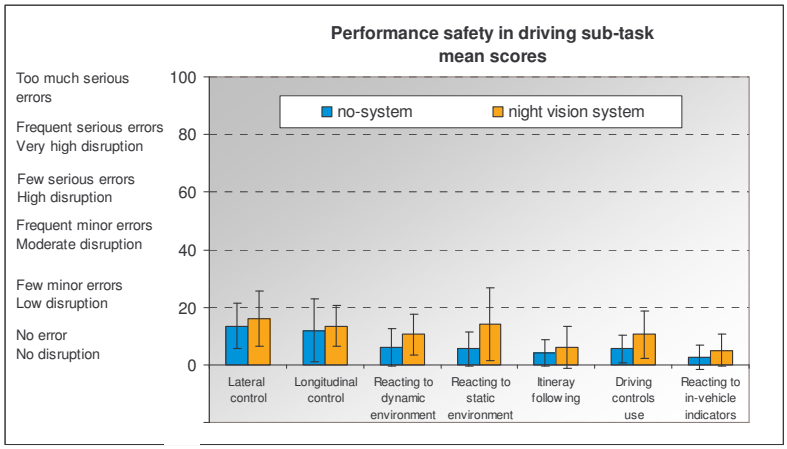



**Fig. 2-17 : PSA-TLX - Performance safety in driving sub-tasks**

Learn to use a night vision system induces low effort in driving as mean scores show very rare difficulties in driving (see Fig. 2-18)

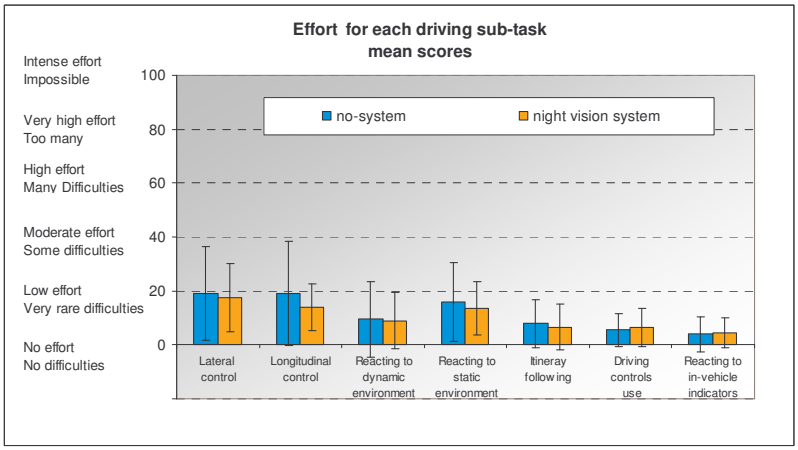



**Fig. 2-18 : PSA-TLX - Effort in driving sub-tasks**

However, variability between drivers is also important. No significant difference appear between no-system driving and driving with the system
Wilcoxon test does not show any significant difference between the condition of driving with no-system and the condition of driving with the night vision system.

The compromise between effort and performance shows that globally, compromise achieved by drivers in driving is efficient as the compromise is favourable even optimal when drivers learn to use a night vision system. Very little unfavourable compromise is reported (See Fig. 2-19)

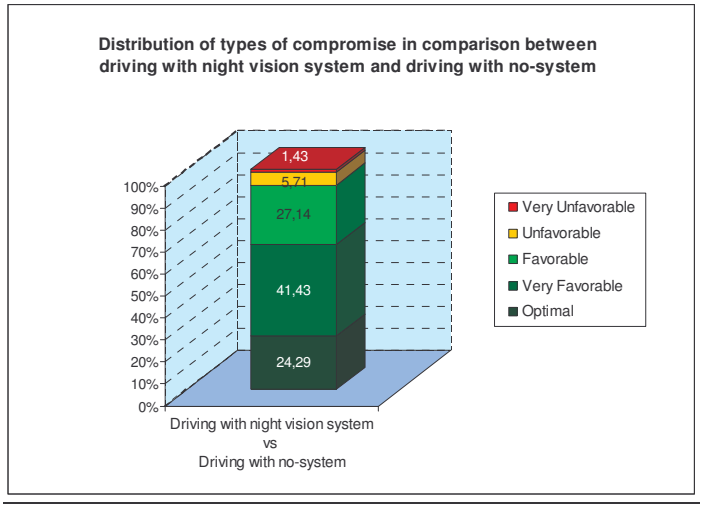



**Fig. 2-19 : PSA-TLX - Frequency of different types of compromise (night vision system)**

Individual analysis emphasizes differences between drivers (see Fig. 2-20)
4/10 drivers have operated a unfavourable compromise between supplied effort and reached performance when driving with the night vision system, this unfavourable compromise only concerns one or two components of driving tasks.
Among these four drivers, one driver reported a very unfavourable compromise in one sub-task of driving when driving with the night vision system.

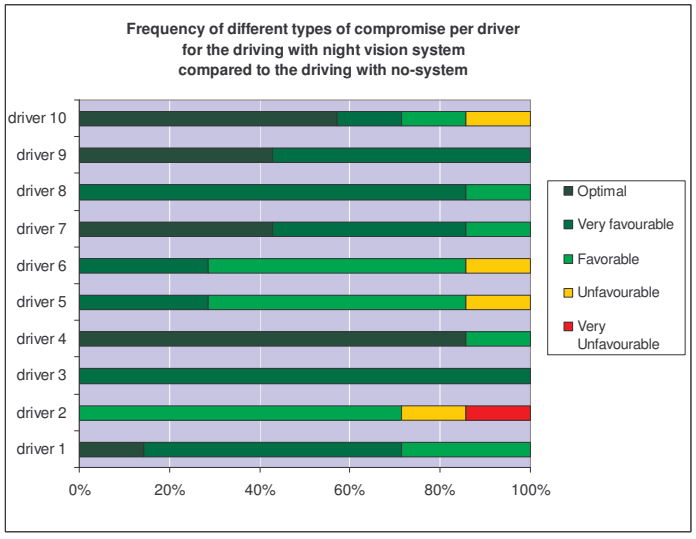



**Fig. 2-20 : PSA-TLX - Frequency of different types of compromise per driver (night vision system)**

Learn to use the night vision system imposes notable driving workload for some drivers: compromise between effort supplied and reached performance, is operated differently with the night vision system, as the driving task's demands really change.

- IVIS : dialling and destination selection (vocal input/output)

PSA-TLX has been tested with IVIS (see Chin E. and Nathan F., 2005).
During the road experiment, 24 young drivers were firstly asked to drive without any secondary tasks, then to use two versions of an embedded telematic system, and to achieve two IVIS tasks: dialling and destination selection. IVIS tasks demands are auditory and temporal as input and output are auditory through a voice recognition system (no visual display).
Depending on the version, the performance of voice recognition and task's demands vary: version n°1 presents high reliability and no correction of system recognition errors is required; while version n°2 with low reliability requires driver's correction of recognition errors in some keywords

Group's trend shows that driving performance is good and acceptable in all sub-tasks: drivers report very low driving disruption. However, variability between drivers is important, especially the performance in taking into account road users.
ANOVA indicates one significant difference between no-system driving and driving with an IVIS requiring recognition errors: performance in taking into account other road users is degraded when using IVIS with low reliable voice recognition, even the average level of performance is acceptable (see Table 2-12)

| ANOVA Performance | Lateral control | Longitudinal control | Reactivity to dynamic environment | Reactivity to static environment | Itinerary following | Use of controls and equipment | Use of safety indicators |
|---|---|---|---|---|---|---|---|
| No-system (1) vs IVIS no errors (2) | NS | NS | NS | NS | NS | NS | NS |
| No-system (1) vs IVIS with errors (3) | NS | NS | **S (p<.03)** | NS | NS | NS | NS |
| IVIS no errors (2) vs IVIS with errors (3) | NS | NS | NS | NS | NS | NS | NS |

**Table 2-12 : PSA-TLX - ANOVA on driving performance**

Mean scores indicate low effort and very rare difficulties met by drivers in all conditions.
However, variability between drivers is important.
ANOVA shows significant differences between conditions for three sub-tasks: in effort supplied to take into account road environment (infrastructure and road users) and to use driving controls. (see Table 2-13)

| ANOVA Effort | Lateral control | Longitudinal control | Reactivity to dynamic environment | Reactivity to static environment | Itinerary following | Use of controls and equipment | Use of safety indicators |
|---|---|---|---|---|---|---|---|
| No-system (1) vs IVIS no errors (2) | NS | NS | **S (p<.004)** | **S (p<.02)** | NS | **S (p<.02)** | NS |
| No-system (1) vs IVIS with errors (3) | NS | NS | **S (p<.005)** | NS | NS | **S (p<.05)** | NS |
| IVIS no errors (2) vs IVIS with errors (3) | NS | NS | NS | NS | NS | NS | NS |

**Table 2-13 : PSA-TLX - ANOVA on driving effort**

When interacting with IVIS, whatever the reliability of the voice recognition (both versions, with or without recognition errors correction), drivers report higher effort in reacting to road

users and to use driving controls. When using IVIS and no correction of system errors is required, drivers mention more frequent difficulties to take into account infrastructure and road signs.
No significant difference is reported in ANOVA between both versions of IVIS in terms of performance safety or effort, as version 1 is used before version 2 (conditions' order is imposed for all drivers), the absence of an effect of task's difficulty induced by recognition errors correction, may be due to a learning effect even if an effect of the system's version (reliability of voice recognition) may exist.

Globally, no major changes in compromise between supplied effort and reached performance, is observed between IVIS whatever the version of the system, and the condition without any IVIS task; as it is between both versions of IVIS (see Fig. 2-21).

Few drivers report unfavourable compromise operated between supplied effort and reached performance in driving activity, almost a half of drivers mention optimal compromise revealing low driving workload when using the IVIS while driving, the other half report favourable even very favourable compromise.

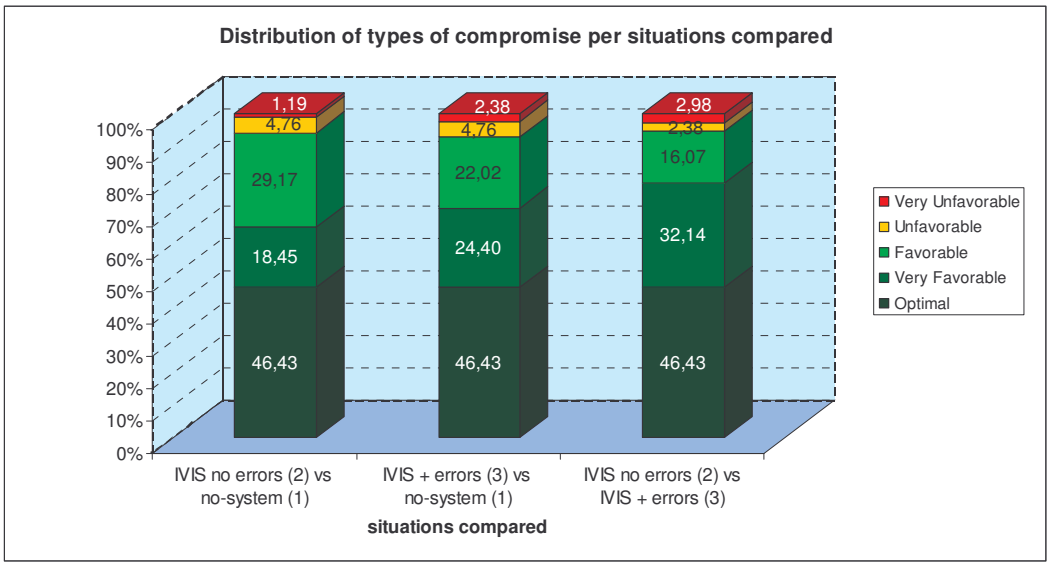



**Fig. 2-21: PSA-TLX - Frequency of types of compromises per situations compared (IVIS)**

## *2.5 CONCLUSION of state of the art*

This state of the art aimed at presenting a rapid and concise review of a number of methods for assessing workload based on subjective judgement.

In total 22 methods (including the last three ones, that were chosen for these trials) are reviewed, namely: NASA Task Load Index, Subjective Workload Dominance, Subjective Workload Assessment Technique, Cooper-Harper Scale and Modified Cooper-Harper Scale, Defence Research Agency Workload Scale, Instantaneous Self Assessment, Bedford Scale, Malvern Capacity Estimate, Performance and Usability Modelling in ATM, Wincrew, Analytical Hierarchy Process, Crew Status Survey, Dynamic Workload Scale, Equal-Appearing Intervals, Flight Workload Questionnaire, Hart and Hauser Rating Scale, Honeywell Cooper-Harper Rating Scale, Magnitude Estimation and NASA Bipolar Rating Scale, and the last ones Behavioural Markers of Drivers Mental Workload, Driving Activity Load Index, and PSA-TLX.
As the objective of this review was to describe the state of the art in this area, no critical review or comparison of methods was carried out. However, it must be noted that certain methods, namely NASA-TLX, MCH, AHP, are widely popular and certainly much more frequently applied than other techniques.

The subjective methods here chosen and presented come mainly from aviation domain. This does not mean for sure that aviation is the only research field in which such assessment activities took place, but surely it is one of the most fruitful. In fact, besides some "traditional" methods that we can find in other research areas, in aviation several other methodologies have been invented, implemented and validated, due to the particular high safety level requested in this domain not only from the people directly involved (the pilot) but also from his "surrounding interfacing elements" (the control tower, the flying crew and the maintenance team). Workload becomes a key issue in performance assessment and its features are discussed during training activities.

As for many concepts and tools studied firstly in aviation domain and then successively exploited in driving research activities (sometimes integrally, sometimes with proper differences taken in consideration), workload in automotive domain have been wisely investigated, using some of the methods presented above (and surely some others, like OW, Overall Workload, here absent).

Of the first 19 methods dealt in this document, 6 have been integrally applied in driving performance evaluation (references texts are 3 official deliverables: RoadSense D 2.1 "State of the art on HMI metrics and target values", AIDE D2.2.1 "Review of existing techniques and metrics for IVIS and ADAS assessment", and AIDE D2.1.1 "Review of existing Tools and Methods"), collecting data about drivers' performance and about the tool itself.

These methods are:

- NASA TLX,
- SWORD,
- SWAT,
- MCH,
- Bedford scale
- NASA Bipolar Rating Scale

They represent the best-established and the most utilized tools in driver workload assessment. Their goodness is inferred by the analysis of some critical parameters, namely Reliability, Validity, Sensitivity, Diagnosticity, Practicality (ease of use), Acceptability, Intrusiveness, Selectivity, Transferability . Many different comparisons between tools have been carried out considering these "classical" criteria, although the results found with some specific subjects could not be confirmed with other subjects (for example, a tool could be diagnostic with pilots, but maybe not with drivers because they can't differentiate two or more factors), because of the domain specificity.

Below we report integrally three summary tables (see Table 2-14, Table 2-15, Table 2-16) , freely taken from a document of the European Organisation for the Safety of Air Navigation (2003) for the INTEGRA project, showing critical evaluations of 3 of the aforementioned tools tested in air traffic management (ATM) studies:

| | |
|---|---|
| **Reliability** | Dependent on operator acceptance; generally high reliability |
| **Validity** | Highly correlated with task performance and other workload measures |
| **Sensitivity** | Sensitive to variations in task difficulty. Monotonic relationship with loading level |
| **Diagnosticity** | Not diagnostic |
| **Practicality** | Easily administered. Decision tree-based, but tree only initially consulted. Subject provides a rating between 1 and 10 |
| **Intrusiveness** | Low intrusiveness |
| **Summary** | Useful as a simple measure of overall workload, but not diagnostic |

**Table 2-14 : MCH Scale evaluation**

| | |
|---|---|
| **Reliability** | Reliable, even if reporting delayed by up to 30 min |
| **Validity** | Underlying dimensions not empirically validated |
| **Sensitivity** | Generally less sensitive than TLX, but still high and extensively tested |
| **Diagnosticity** | Multidimensional: three scales (time load, mental effort load, psychological stress load). These have been found to be differentially diagnostic (Moroney, p.101) |
| **Practicality** | Two steps: scale development (scale ranking of 27 scale combinations); event scoring. |
| **Intrusiveness** | Slightly more demanding than other subjective measures (Step 1 can take up to 45 min (Wierwille & Eggemeier, p.268). Lower user acceptance than TLX. Completed off-line |
| **Summary** | A test that provides useful data, but requires more effort than some other multi-dimensional scales and therefore not recommended in the present context |

**Table 2-15 : SWAT evaluation**

| | |
|---|---|
| **Reliability** | High reliability |
| **Validity** | Extensively validated |
| **Sensitivity** | High sensitivity |
| **Diagnosticity** | Multidimensional: six scales (Mental, Physical, Temporal, Effort, Performance, Frustration); differentially diagnostic; global score can also be calculated |
| **Practicality** | Two steps: event scoring and paired comparison weighting process to determine the importance of each factor for the task in question (the latter not necessarily required) |
| **Intrusiveness** | Takes 1 or 2 minutes to complete, off-line |
| **Summary** | A very well-established test with a sound basis, which could be used in INTEGRA studies |

**Table 2-16 : NASA-TLX evaluation**

Of all the criteria introduced above, only six are considered in this table, ignoring Transferability, that recover in this work a key role. If we want to bring aviation experience in automotive domain we have to keep in mind the problem of transferability: the terminology and the parameters evaluated by each tool must be tailored to driver reality (being the above criteria values modifiable).

Transferability is one of the evaluation criteria often reported in literature reviews (as in AIDE Deliverable 2.1.1 “Review of existing Tools and Methods” by CRF). Tools and techniques are compared and their goodness is inferred taking into consideration also their transferability to other situations/domains.

In this work, whose aim was to report the state of the art of subjective methods providing an input to work already done in other domains like aviation, the question of Transferability is difficult to address and maybe not so useful in this phase of the project. It has to be remembered that only some experimental sessions or real road test experiments could suggest how a particular tool is functional (therefore valid) and transferable to automotive domain, with different terminology used and different target subjects.

Part of the tools presented in this document have been already discussed in the aforementioned AIDE Deliverable and in another more specific review written by VTEC inside the project, which is Deliverable 2.2.1 “Review of existing techniques and metrics for IVIS and ADAS assessment”. They are here reported again as recurrent tools used in aviation, while other questionnaires have not been reported although in some cases they are a good example of adaptation to a different domain of a previous famous tool, like DALI by INRETS (which is a revision of the NASA-TLX), or the PSA-TLX, a multi-dimensional assessment questionnaire from PEUGEOT-CITROEN.

As pointed out in other review works (RoadSense D2.1) every tool has strength points and drawbacks. Considering both positive and negative aspects and the transferability key issue, we can suggest that, of the methods presented in this work, SWAT could have been exploited in AIDE project. In this tool workload is assessed by three dimensions (Time load, Mental effort load, Psychological Stress load) that could be utilised and easily distinguished by drivers. Tasks while driving are different in terms of attention to be paid: there are mechanical operations as well as tactical and strategic ones (Allen, Lunenfeld, & Alexander, 1971). These 3 levels can be linked to the skill-rule-knowledge conceptualization held by Rasmussen (1987) as it has been done by Peugeot-PSA for realizing the PSA-TLX, and they can be helpful for a further better definition of the 3 variables (with the intent of improving diagnosticity). Different studies have in fact showed a not complete independence of the

variables, because ratings tend to increase for all three dimensions even if only one dimension is altered.

The 3 levels for each variable could be improved, although this could generate a combinatory explosion of conditions to be assessed in the first phase of the procedure. In fact, length and complexity of its administration are weakness points of this method. But the preliminary ranking phase permits to construct personal scale for every subject, therefore to account for individual differences. Being AIDE a project for realizing an adaptive interface, different workload values from different drivers in the same situation are surely useful data.

Moreover, there are some alternative versions of this tool, that substituted the preliminary card sorting task with a pair-wise comparison one, allowing to spare time and to avoid some possible errors. Another alternative version used a continuous rating scale instead of a discrete one.

What it is important to realise is that most of the methods have been developed with a reference to specific domains of application. Their expansion or implementation to another domain may require some careful adaptation and revision before enabling full applicability and validation. Experience brought from the aviation system could be fruitful in suggesting some starting points for workload assessment, but sure not in providing some universal tools, because of the important differences between a driver (a common person who choices its vehicle and who drives for many different reasons, and being differently motivated) and a pilot (a trained expert person who pilots only some specific aircraft and who is working when piloting).

# 3 Input from previous AIDE work

A review of existing techniques and metrics for IVIS and ADAS assessment was carried out in AIDE, task 2.2.1 (Johansson et al., 2004). One chapter in this review focused specifically on subjective assessment methods; the experiments in this task are partly based on what was discussed in the earlier review. In D 2.2.1, a first short description of PSA-TLX and DALI can be found among others methods and tools.

In state-of-the-art of previous section, some tools are described again besides a set of tools specifically taken from aeronautics that show how "workload" construct has been studied and assessed with very specific instruments.

In D2.2.1, for every tool presented, some strong points have been introduced. The experiments carried out in this work have partly confirmed them. Regarding PSA-TLX for example, it was argued that it was "better than NASA-TLX in driver's understanding of scale terminology": this was one important outcome of experimental trial in Task 2.2.6, as subject indicated it as a tool made of clear and understandable concepts (as reported also in Section 8, below).

Regarding multi-dimensional scales, in D 2.2.1 there was a section including some issues for further research (especially for DALI),. Trials carried out in this work have represented of course an important validation opportunity, that in D 2.2.1 was indicated as still missing.

Workload have been dealt also in another deliverable of AIDE Project, D 2.1.1 "Review of existing Tools and Methods", written by CRF. Among different methodologies of workload assessment (performance measures, physiological measures as well as expert-reported measures) also subjective measures were presented and PSA-TLX and DALI have been introduced. One of the aims of this great review work was to presents different methods of assessment (not only subjective as the state-of-the-art in D 2.2.6) of different important constructs (not only workload but also Situation Awareness or User Acceptance for example), therefore the description of this tools are quite synthetic and find in D 2.2.6 a more complete presentation.

Workload is a great issue of research that have been treated since AIDE Project started, firstly in a more general manner and then in a more and more detailed one. In this deliverable a further state-of-art was written also in order to show how workload have been investigated in other domains, using specific tools with appropriate terminology. Of course some of this tools could not be adapted to automotive domain, but they can demonstrate that workload can be successfully assessed if our evaluation tools are tailored for the situation, the user and the particular process under evaluation, as well as it has been done for PSA-TLX, DALI, and finally, for BMDMW.

# 4 Contribution to overall AIDE objectives

Results of experiments conducted in task 2.2.6 are compared with main pros/cons identified in task 2.2.1 for PSA-TLX, DALI and BMDMW.

The results from the work in this task have been provided to task 2.2.7 where workload evaluated with subjective methods will be compared to the other metrics and tools developed within Workpackage 2.2. The results from the 2.2.7 work will be incorporated into the generic test regime developed in Workpackage 2.1 and will be used to assess the AIDE demonstrators as well future ADAS and IVIS.

# 5 Experiment 1 (CRF)

The present research was carried out by Centro Ricerche Fiat to test the validity and reliability of three Workload Subjective methods.

In particular, the aims were:

- To test the instruments applicability for different systems (IVIS, ADAS) and for a Virtual Reality scenario;
- To refine the procedure of administration (i.e. improve instructions, modify or improve factors).

As one of the aims of AIDE project is to develop a general methodology to evaluate IVI and ADA systems at different levels of development, it is very important to verify the reliability of subjective workload instruments also in context different from a real road scenario. With these trials, CRF wants to understand the level of applicability of different types of subjective tools in Virtual Reality context, offering useful suggestions to increase their suitability.

At the end of trials CRF, together with other T 2.2.6 Partners, wants to define the "best" subjective tool in terms of sensitivity, diagnosticity, acceptability, duration and complexity of processing.

## *5.1 Method*

The research was carried out in a VR simulator (CAVE© Simulator). Participants interacted with two types of systems:

- an IVI System (radio);
- an ADA system (Frontal Collision Warning system, FCWS).

### 5.1.1 Intended users

As one aim of research was to use the VR Simulator, the sample of participants was chosen from either VR Experts (more than three trials in previous experiments at the simulator) or VR non-experts.

### 5.1.2 Participants

Sixteen participants were recruited for the experiment. Their participation was voluntary.
The sample presents the following characteristics:

- **Driving licence:** 12 years mean (min 5 years, max 19 years);
- **Annual driving kilometres:** 20.125 km mean (min 10.000, max 30.000)
- **Age**: 30 years mean (min 23 years, max 37 years)
- **Gender:** 6 female, 10 male
- **Same level of experience/knowledge** of the ADAS system (no knowledge at all);
- **Experience with Virtual Reality simulator:**
  - 9 experts (more than three experiences in a Virtual Reality simulator);
  - 7 non experts (no experience before this trial).

### 5.1.3 Apparatus

**VR Simulator Description**

CRF VR Simulator allows, with suitable settings, to simulate a realistic vehicle driving.
The system is the result of the integration of the following parts (see Fig. 5-1):

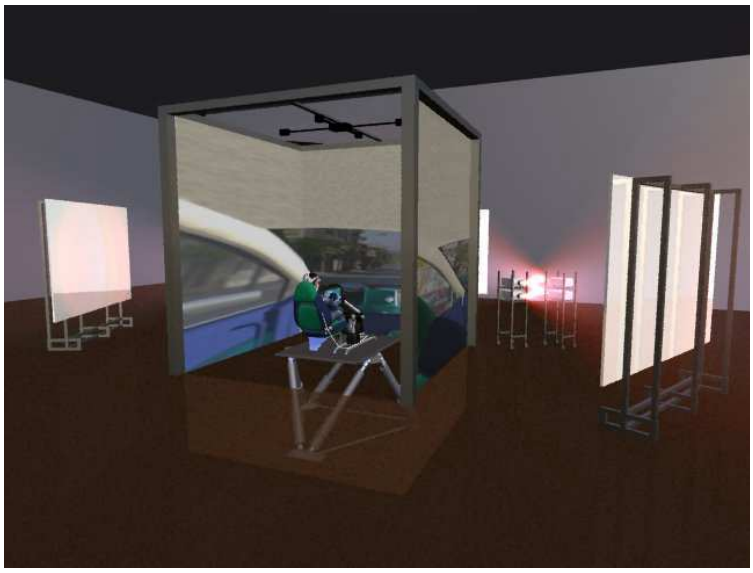


**Fig. 5-1: The CRF Driving simulator architecture**

- An I-Space with:
  - 3 orthogonal rear projected screens, 3m high and 3m wide. They allow, to the observer placed in the ideal cube centre, to enjoy a ±45° vertically, and ±135 horizontally field of view;
  - 6 projectors (2 per screen) which allow to obtain the stereoscopic visualization with a 1280x1024 pixel resolution;
- A 6 degree of freedom mechanical platform with a bandwidth of about 10 Hz;
- A physical mock-up mounted on the mobile platform. This mock-up includes:
  - An automotive seat, adjustable in longitudinal and tilting way;:
  - A steering wheel with force feedback;
  - Brake and accelerator pedals;
  - Automatic or Selespeed gear shift.
- A surrounding audio system, including a subwoofer and 4 tweeters that are mounted on the mock-up;
- A motion tracking system that measures and provides the driver head position within the I-Space;
- A PC cluster that manages the graphics (frame rate: 30÷60 Hz, depending on the interiors and scenario complexity), the dynamic platform and the vehicle dynamics accurate simulation (16 d.o.f).

**The CRF driving simulator: functional description**
The CRF driving simulator, because of its composition, has the following functional characteristics:

- The virtualisation of the vehicle interiors which aren't physically realized in the mock-up, with the possibility to change "on the fly" their configuration;
- The reproducing by means of the dynamic platform of:
  - The vibrations due to the road roughness and to the engine;
  - Car body motions (pitch and roll);
  - Motion cueing that are the inertial forces perceived by the driver during braking/accelerating manoeuvres and/or during a bend.
- The extremely realistic simulation of the vehicle dynamic behaviour, with the possibility to set hundreds of model characteristic parameters, including the road surface adherence coefficient;
- The continuous graphic visualization updating, based on the driver head and platform positions;
- The possibility to utilize, properly developed, the most suitable for the specific experimentation 3D scenarios;
- A vehicular traffic model populating the road scenarios according to the road signs (see Fig. 5-2). Every vehicle is capable to adjust his behaviour according to the others, included the driver's one, and to the experimentation needs to generate useful situation, such as "dangerous" ones;
- The vehicle typical internal noise reproducing, taking in account the 3 main components: engine, aerodynamics and tyres rolling. Moreover, it is possible to reproduce external acoustic sources, according to the sound wave propagation rules (propagation velocity, atmospheric absorption, Doppler effect, etc.).

**Fig. 5-2: Examples of scenarios populated by vehicular traffic**

Functional characteristics are always improving, allowing the CRF Drive Simulator to operate in the following main fields:

- The new ADAS and NIT devices acceptability and usability analysis;
- The study of new force feedback logics applied to the primary controls, or the study of new primary controls typologies (joystick, cloche, etc.);
- The dynamic-vibrational comfort analysis, through the simulation of several suspension types;
- The evaluation of an interior extern visibility (direct or through rear view mirrors).

**VR Scenario**
The VR simulator scenario allows to obtain a **controlled environmental situation**. In this way, different and equal level of situation complexity among participants could be defined. All participants drove in the same simulated scenario with different situations (traffic agents performing merging, overtaking, suddenly lane changing and sudden braking) which forced participants to do various driving maneuvers.

**Type of road**
Eighteen kilometers of simulated highway with right and left curves. The highway has three lanes with one emergency lane. Different situations have been created in order to test the functionality of Frontal Collision Warning System:

- Roadway bottleneck;
- Leading vehicles suddenly braking at several instants and situations;
- Stationary vehicle in the middle of the lane.

**Fig. 5-3: Simulated Scenario, shrinkage with static signs**

**Weather condition**
Various fog banks which create different levels of visibility. Three different levels of Visibility can be identified:

- Good visibility: visibility up to 150 metres;
- Medium visibility: visibility up to 60 metres;
- Bad visibility: visibility up to 30 metres.

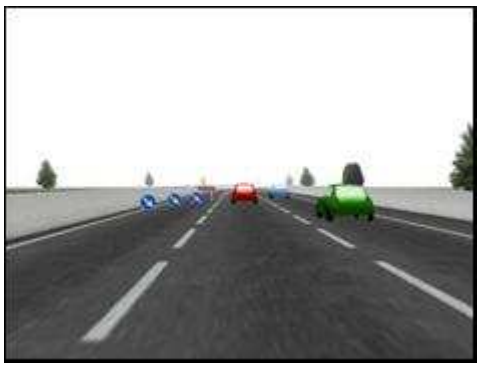


**Fig. 5-4: Simulated Scenario, another view of shrinkage**

**Traffic conditions**
Medium level of traffic flow with intelligent vehicles doing different manoeuvres. Vehicles which suddenly cut off the street to the participant vehicle; vehicles with a very low velocity.

**Driving task**
To avoid any additional artificial workload, drivers are asked to adopt their spontaneous speed and drive as natural as possible.

**Path Definition**
The Path was divided in different *segments*, according to events taking place during the trials.

In particular:

- **Normal** or **Task**: seven path segments in which no events or Task events took place according to the experimental Condition;
- **Brake**: seven path segments in which the leading vehicle suddenly brakes at random instants;
- **Obstacle**: path segments in which a vehicle suddenly cuts off the lane to participants' vehicle;
- **Shrinkage:** path segment in which a signed roadway bottleneck is present;
- **Unexpected action:** path segment in which a stationary vehicle in the middle of the lane is present.

Segments are put together according to an order that alternates Normal/Brake segments (No consecutive instances of Normal segments or Brake segments) and other types of Segments. For example, the first seven segments are formed by a Normal/Task segment followed by a Brake segment, that is followed by another Normal/Task segment and so on. After this sequence, The Obstacle segment takes place. After the Obstacle segment, another Brake segment takes place, as the segment before the Obstacle segment was Normal/Task segment.

Segments apart from Shrinkage alternate three different Fog Banks, in order to provide different conditions to use ADAS system. Participants, therefore, found themselves in dense fog banks when they were requested to operate with IVIS.

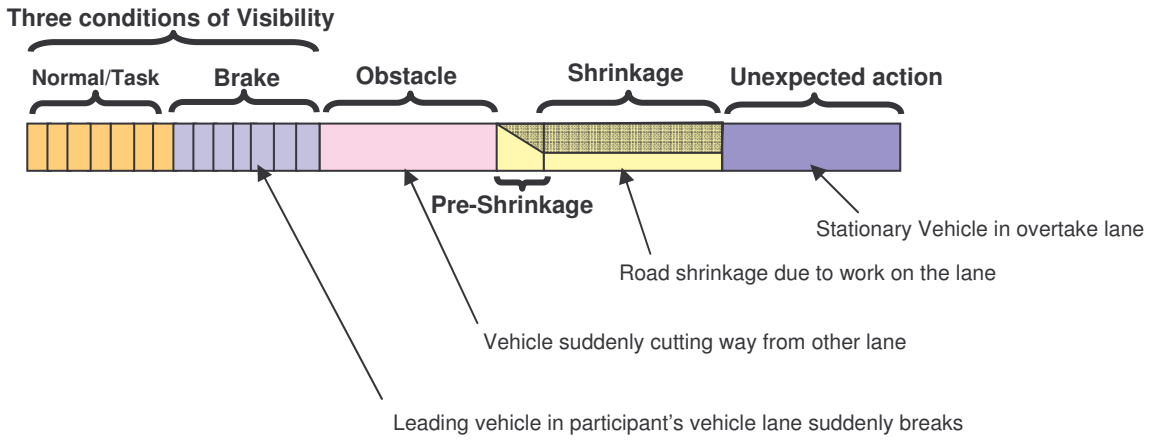



**Fig. 5-5: Types of Segments in VR Trial**

### 5.1.4 Data analysis

**Subjective instruments**
The three subjective workload instruments chosen, are briefly described in the following:

- **BMDMW (BEHAVIORAL MARKERS OF DRIVER MENTAL WORKLOAD)** (see section 2.4.1)
- **DALI (Driving Activity Load Index) developed by INRETS/LESCOT** (See section 2.4.2)
- **PSA-TLX (Task Load index) developed by PSA research department,** (see section 2.4.3)
- **Workload questionnaires' evaluation: final questionnaire**

A specific questionnaire has been designed to get the drivers' point of view on PSA-TLX and DALI (see independent appendix IV.2)

10 questions are asked to drivers dealing with :
1. questionnaire presentation
2. time required to fill in
3. scoring difficulty
4. questions understanding
5. concept understanding
6. written explanation understanding
7. oral explanation understanding (necessary/sufficient)
8. effort required to fill in
9. evolution of effort along times
10. exhaustivity

These questions are asked for each method. Then 6 other questions dealing with global feeling, comments and suggestions are collected.

### 5.1.5 Objective Data analysis

Trials have been conducted using a VR Simulator. This approach allowed a high control of all experimental conditions (repeatability of trials in nearly the same conditions), as well as the possibility to record all *objective* data (speed, brake pedal action, lateral and longitudinal actions, other vehicles on the road, duration of Warnings, ADAS outputs etc). No secondary task value has been recorded, but the effect of IVIS was studied in terms of variations in levels of Demand during the trial (how much does the interaction with IVIS augment the Demand along the trial?).
The analysis of collected data provides a medium to assess the efficacy of the three questionnaires regarding their ability to "capture" the different demands of the three driving sessions.
The experimental design has a variability due to different confidence levels with drive and simulator, very complex tracks with vehicles, curves, low visibility and hard tasks, as well as

interacting variables. In order to clear such variability, a procedure to calculate *variances* from a *baseline performance* is required.

A "**baseline**", using the condition "No Systems", has been calculated. Only the segments without events have been taken under consideration. As three different Visibility levels might show, three different baselines (correspondent to three visibility levels) have been computed.

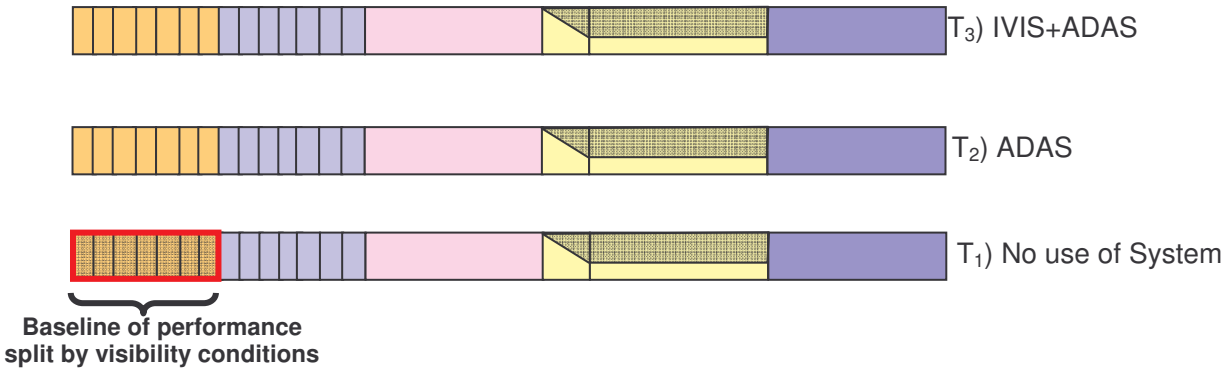



**Fig. 5-6: General scheme of the three experimental conditions**

In Condition $T_1$ no events happened, whereas in condition $T_3$ participants were requested to execute a difficult task with an IVIS while the ADA System was turned ON.

**Metrics**
Objective Data recorded has been analyzed in order to identify values which could give information about the variations in Demand during the trial. Each participant drove in the three Conditions, so data about a particular segment in a particular Condition of a particular participant was at disposal to make data analysis.

**Metrics Calculation**
Metrics defined for a segment were computed as the Standard Deviation of values related to the Metric.
Some metrics have been split according to the Visibility Level. As a result, three values for every metric, each dealing with a different level of Visibility, have been calculated.
It is therefore possible to inquire the effect of different Visibility on various aspects of driving

**Metrics Definition**
Particular *metrics* have been identified to assess Demand in trials. They are:

- Mean Speed;
- Number of accidents;
- Mean duration of Warnings (Red/Yellow) : time during which warning signals are on;
- Braking reaction timings: two measures are set down:
  - time between Warning switching on and brake pedal pressure;
  - relative distance from the obstacle in the moment the brake is activated
- Mean action on brake: Brake pressure angle;
- Headway;
- Relative distance from closest vehicle (when present);
- Lateral displacement from the centre of lane.

The most representative metrics for each type of segment are listed in the following table:

| Segment | Metrics | Segment | Metrics |
|---|---|---|---|
| **Normal/ Task/ Obstacle** | - Mean Speed<br>- Number of accidents<br>- Warning duration (red/yellow)<br>- Brake reaction time<br>- Action on brake<br>- Headway<br>- Relative Distance<br>- Lateral displacement | **Shrinkage** | *During shrinking*<br>- Mean Speed<br>- Brake reaction time<br>- Action on brake<br>*While in car-following*<br>- Warning duration (red/yellow)<br>- Brake reaction time<br>- Action on brake<br>- Headway<br>- Relative Distance |
| **Brake** | - Warning duration (red/yellow)<br>- Brake reaction time<br>- Action on brake | **Unexpected Action** | - Warning duration (red/yellow)<br>- Brake reaction time<br>- Action on brake |

**Table 5-1: Summary of Metrics in Objective Data Analysis**

**User Demand Index**

The behaviour of user is not predictable, so it is not possible to rate metrics in order of importance, because such order is not universal, and different users can give more importance to different aspects monitored by the metrics; for instance, a user can give more importance to maintain a regular longitudinal speed, in spite of lateral control; another user may show a reverse behaviour. It is therefore important to take into account all influences of performance due to user behaviour.
This problem requires the identification of an *Index of User Demand* during the trial, representing its *overall demand* as a result of its behaviour as observed through the metrics of a specific segment.

The process of index creation is the following:

1) Standard Deviations of Metrics, at the level of a Segment in a Condition, are calculated;
2) If a metric for a specific segment shows a significant difference across conditions, it will be normalized (z-points), hereby representing a specific index of variability for the metric;
3) Normalized metrics showing differences are then summed together, and the result will be the *Synthetic Index of Variability* in the Segment for a given Condition of a particular participant;
4) Global Metrics are considered as significantly different if one of the following holds true:
    a. The metric is globally different across conditions, or
    b. It is locally significantly different in any segment in the corresponding condition.

5) Global Metrics showing differences are summed together, and the result will be the *Global Index of Variability* for a given Condition of a particular participant.

It is important to remark that significantly different metrics must be normalized *before* calculating the index. The reason is that metrics do not have the same reference system; for example, Mean Speed is computed in Kilometers per hour, whereas Brake pedal Reaction time is computed in seconds. The process of normalization simply maps all distributions in a Normal Distribution with mean 0 and SD ±1, thus enabling metrics to be comparable.

**Normalization**

An example will be useful to better understand the normalization process. Let us consider Global Mean speed, whose dispersion is depicted in Fig. 5-7:

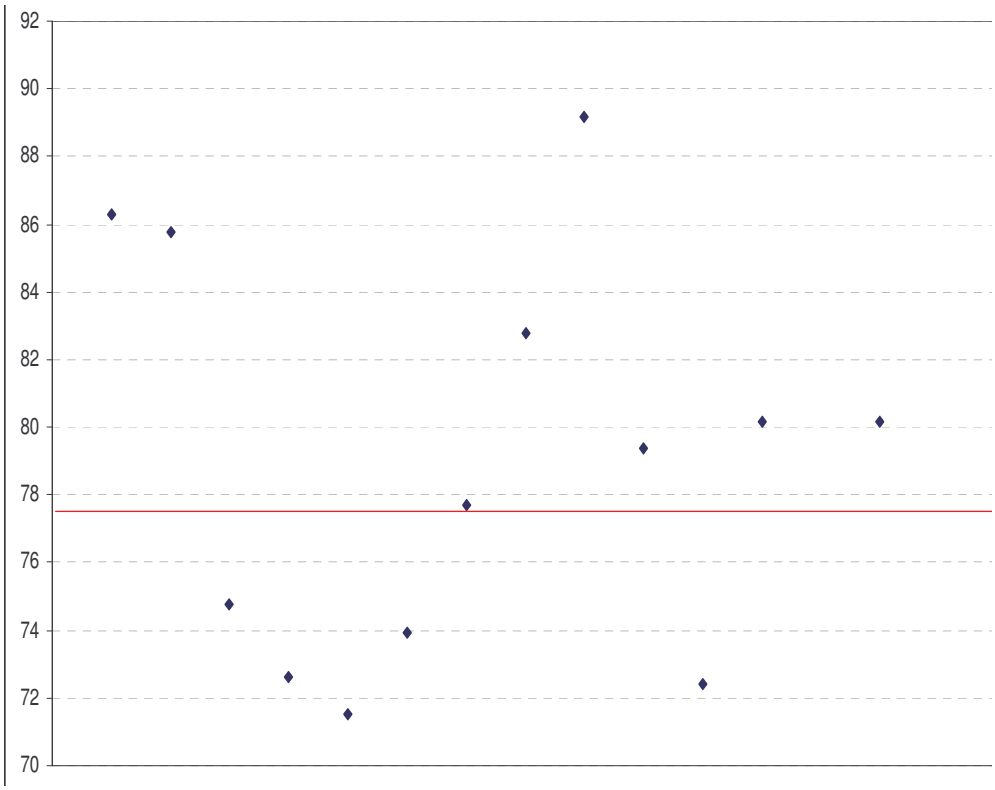



**Fig. 5-7: Global Mean Speed Dispersion**

The red line indicates the Global Mean Speed, at about 77 kms/h.

Square Errors Distribution is then computed, as depicted in Fig. 5-8:

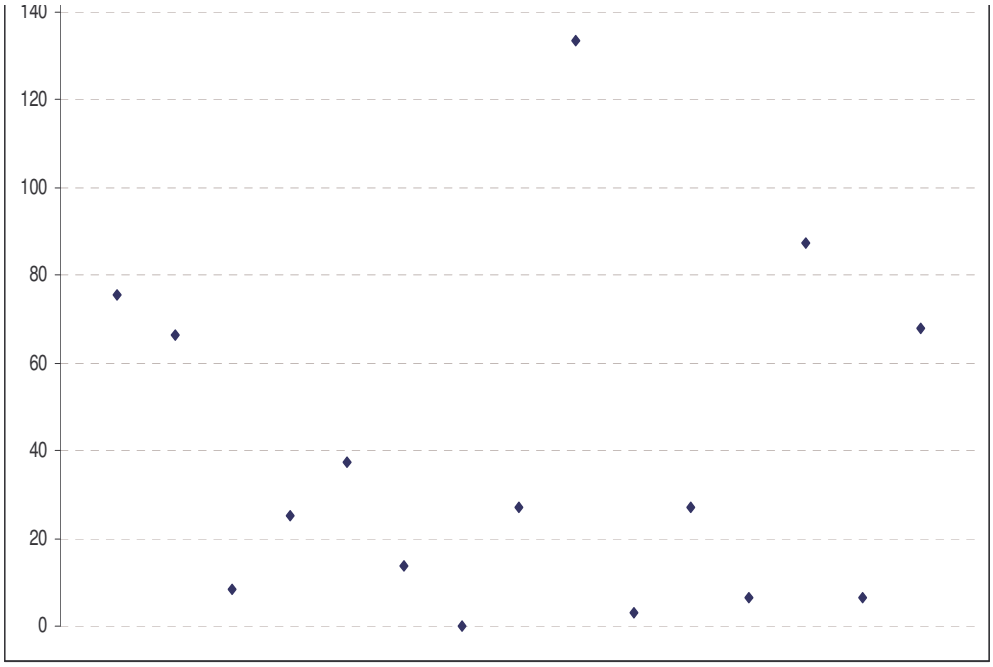



**Fig. 5-8: Square Errors Distribution**

The scale is still in Kilometres per hour, and therefore the distribution can be used conjointly only with other distributions with the same scale.

The last step is the normalization of the Square Errors distribution, giving the distribution depicted in Fig. 5-9:

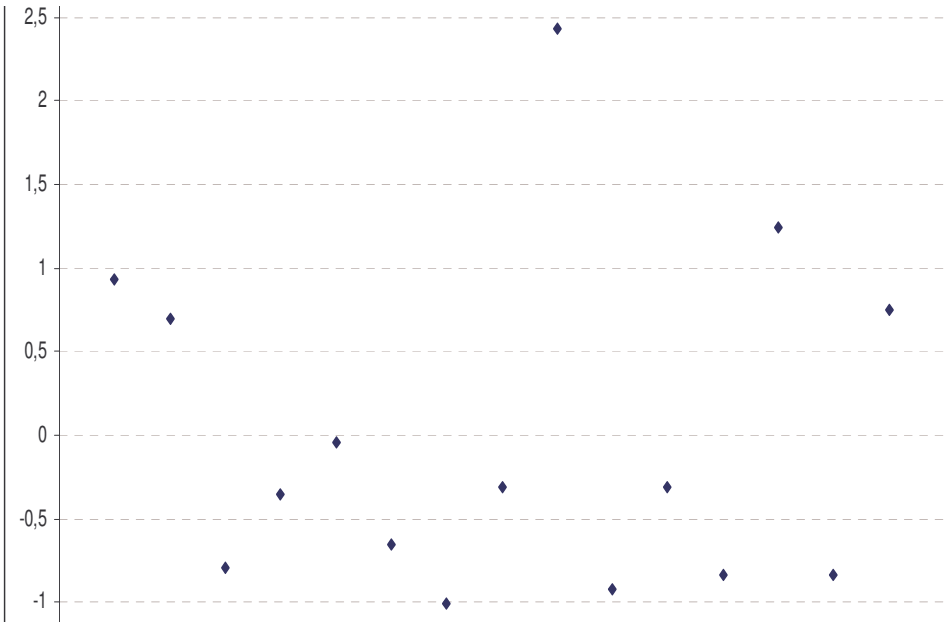



**Fig. 5-9: Square Errors Normalized Distribution**

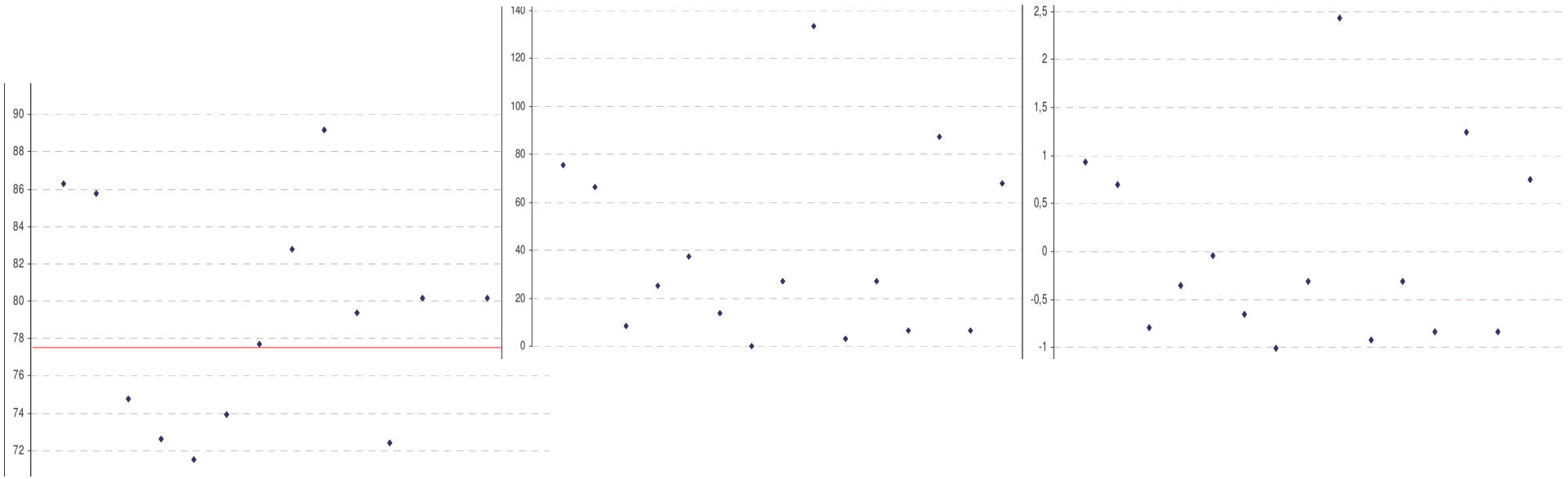



**Fig. 5-10: Relationship among Distributions**

**Relationship among distributions**

Starting from Fig. 5-10b, we see that lower values are closer to the mean indicated in red in Fig. 5-10a (the deviation is lower). So, the origin of Y-axis of picture Fig. 5-10b is coincident with the Mean in Fig. 5-10a, with the meaning that a value of zero for Square Error means coincidence with the mean. But, as distribution in Fig. 5-10c is the same as distribution in Fig. 5-10b as can be easily seen via confrontation of the graphics, the lowest value in distribution in Fig. 5-10c must coincide with the lowest value in Fig. 5-10b. This leads to the arrangement of distribution depicted above.

Scales of axes in graphics are not the same, but the focal point is the origin of data, and not their magnitude (at least not in this case), as we are interested in explaining why lower values in the Normalized distribution are better.

Original SD Distribution of Mean Speed and Normalized Distribution of Mean speed can be now directly compared:

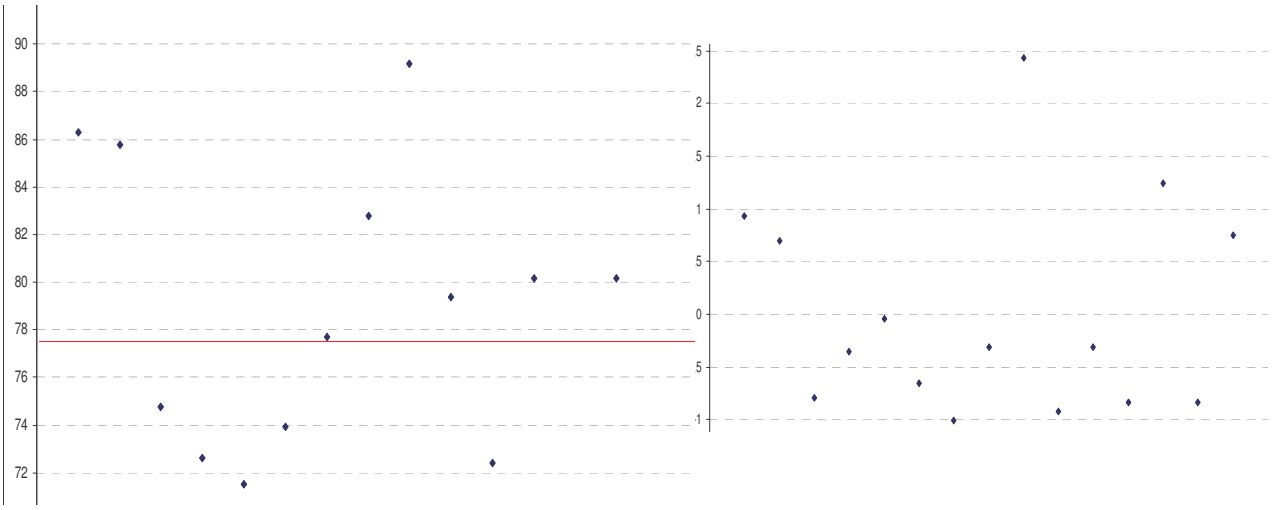



**Fig. 5-11: Comparison between Speed Distribution and Normalized Speed Dispersion**

From the picture above, the following results can be achieved:

- **Lower values in Normalized distribution are closer to the original mean** (thus being *better* as they are less variable), whereas
- **Higher values in Normalized distribution are farther from the original mean** (thus being *worse* as they are more variable).
- **Summation** of single Metrics indices is possible, giving an overall index of a particular region of the track (either a segment or the whole track). As some values might be negative (better) or positive (worse), all contributes from relevant metrics are taken into account to create a single value, to be interpreted as outlined above.

Global Indices can be used to obtain a quick but accurate trend of Demand along all the track across Conditions; Segment Synthetic Indices can be used to study the trend of Demand in different parts of the track, to better understand where variations were more consistent.

**Computation**
As an explicative conclusion, the Indices of Brake Segment for Condition 1 (No Systems) and Global Index for Condition 1 are computed as follows:

$$IBrake(T_1) = N(YellowWarning(T_1)) + N(RedWarning(T_1)) + N(BrakeReactionTime(T_1)) + N(NActionOnBrake(T_1))$$

$$I_{T1} = N(MeanSpeed(T_1)) + N(NumberOfAccidents(T_1)) + N(YellowWarning(T_1)) + N(RedWarning(T_1)) + N(BrakeReactionTime(T_1)) + N(ActionOnBrake(T1)) + N(Headway\ T_1)) + N(RelativeDistance(T_1))$$

*where*

*Is* : *Index* of Segment or Condition *s*;
*Distribution* ∈ *{MeanSpeed, YellowWarning, RedWarning, ActionOnBrake…}*

*N(Distribution(Tx)) :* Normalized *Distribution* Value in Condition $T_x$ *(x = 1,2,3)* if *Distribution* across conditions is significally different, zero otherwise

**Deeper analysis of secondary task impact**

In order to better understand the real impact of secondary task in IVIS+ADAS condition, a deeper analysis using metrics sensitive to secondary task is done. The track segments taken in consideration for this analysis are only the segments where an interaction with the IVIS takes place, and the corresponding segments in the other two conditions.

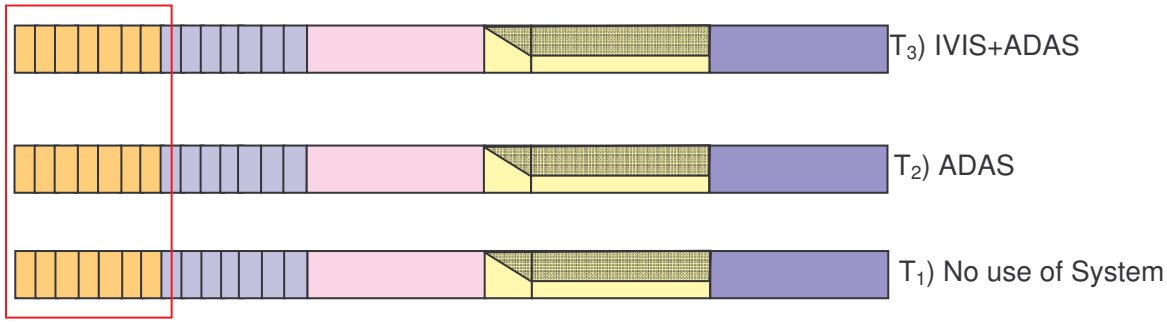



**Fig. 5-12: Segments taken into account in specific secondary task analysis**

The two metrics used in the deeper analysis of secondary task impact are:

- Lateral Displacement
- Mean speed

Objective results will be presented only in terms of significances in differences of Mean Speed Standard Deviation and Lateral Displacement Standard Deviation.

### 5.1.6 Tasks/Systems (e.g. IVIS tasks or ADAS systems, both…)

An ADA System, a Frontal Collision Warning System (FCW), has been used. The FCW warns participants either by visual and auditory means when they are approaching too fast to leading vehicles. Either current relative speed from leading vehicle and relative distance are taken into account to determine whether a warning should be triggered. If a warning is to be triggered, it can be:

- A *yellow* visual light on the front of the dashboard, as well as a *pulsing* sound – this warning means that the leading vehicle has a combination of speed and distance that are not dangerous for the driver when current driver's speed is maintained. However, caution is required as an obstacle is present in our lane;
- A red visual light on the front of the dashboard, as well a more intense pulsing sound – this warning means that, if current driver's speed is not reduced, there is the possibility of an accident with the leading vehicle.

An IVI System, a radio, has been used during the trial. Participants had to perform a Difficult Task with the System while driving:

- Enter the Equalization menu;
- Select a specific option (four steps to reach the Customer option);
- Select the first option, change its level of four points, and
- Select the second option.

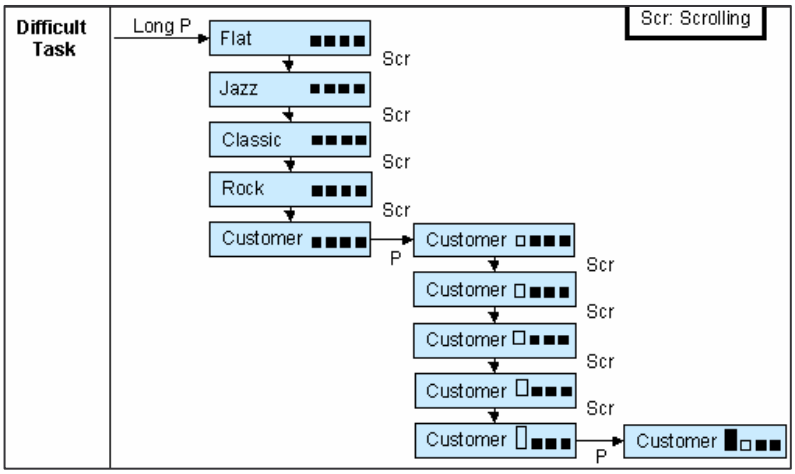



**Fig. 5-13: Difficult Task Scheme with IVIS**

This kind of task has been already tested during ROADSENSE CRF trials.

### 5.1.7 Design and Procedure

A design of one factor (Use of the systems) at three levels was chosen. In the following, a description of the levels of the system is provided.

Use of the systems:

- *No use of systems* (Described in the following as Condition $T_1$) – participants drove without any influence from in-car devices, only relying on their personal background and experience;
- *Use of ADAS* (Described in the following as Condition $T_2$) – participants benefit from the aid of a Frontal Collision Warning (FCW) system.
- *Use of IVIS and ADAS together* (Described in the following as Condition $T_3$) – beside the use of FCW as described in section 5.1.6, participants were asked to carry on a difficult task (already tested during ROADSENSE CRF trials (see Deregibus et al, (2004)).
  Participants repeated the task seven times during the trial.

Hypothesis
On the basis of literature, it has been assumed that the Demand should:

- Go up from the condition "No use of systems" to the condition "use of IVIS and ADAS together".
- Decrease from baseline to ADAS condition, above all in critic driving conditions (low visibility).

The different subjective mental load evaluation methods should be able to discriminate among these different conditions.

Procedure

A pilot test, lasting one day, was done with two subjects; thanks to results of this test, the following procedure was defined.
The experiment was divided in four phases. Each participant familiarized with the simulator in order to lower the bias of unfamiliarity. After the familiarization session, participants began to drive in one of the three Conditions. Conditions were randomized for participants, in order

to eliminate (or lower) the learning effect and distribute it among the three trials. Participants were requested to drive in one of the Conditions for about seven minutes; in Condition $T_3$, participants were requested to interact with the IVIS as described in section 5.1.6. After each trial session, participants were requested to compile one of the three questionnaires. The order of questionnaires was randomized for each participant, again to lower statistical biases.
Each phase ranged in time from 25 up to 50 minutes, comprising driving sessions and questionnaires filling in. Consequently, a global session for each user lasted about 90-120 minutes.

| Phases | Driving | System activation | Time | PSA | DALI | BMDMW |
|---|---|---|---|---|---|---|
| 1 | | VR simulation familiarization | 10’ | | | |
| 2 | ✓ | No use of system | 25’-50’ | ✓ | ✓ | ✓ |
| 3 | ✓ | ADAS | 25’-50’ | ✓ | ✓ | ✓ |
| 4 | ✓ | IVIS and ADAS (difficult task) | 25’-50’ | ✓ | ✓ | ✓ |

**Table 5-2: Experiment Phases and timings**

The experimental trials were conducted during seven days.

## *5.2 Results*

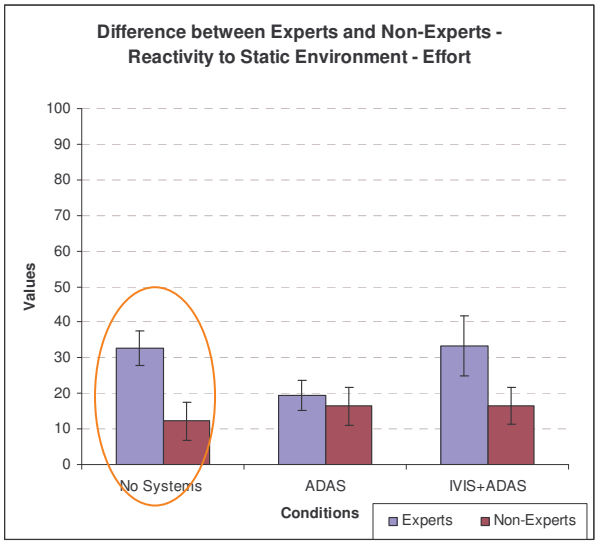
Subjective Results are provided, divided by tool. Objective results are then reported, followed by a comparison between objective and subjective results.

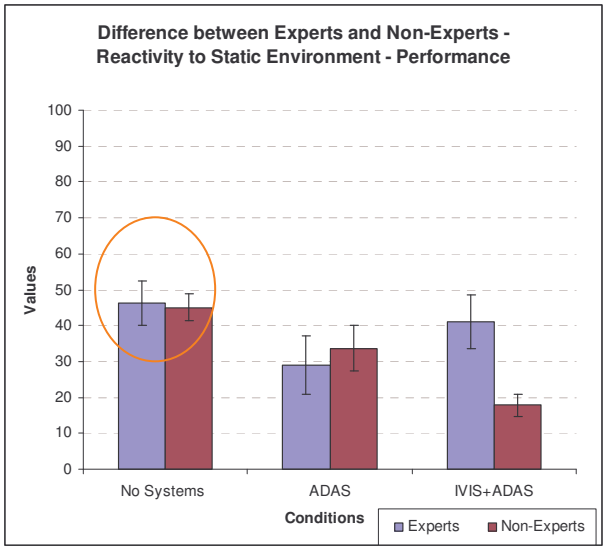
Subjective data collected has been analyzed. Results are reported below, with a description of main considerations. A separate analysis has been done on VR Experts and VR non-Experts, to see whether the experience with the VR Simulator was a factor in Demand variability.

### 5.2.1 PSA-TLX results

**Differences between VR Experts and VR Non-Experts**

A comparison, through non parametric test (Mann-Whitney test), was carried out to verify the differences of judgments between expert and non-expert in VR simulators.
Below, only the significant differences are reported.




| Mann-Whitney Test : Significance | |
|---|---|
| No Systems | *Static Environment*<br>Performance: p=0,042, Z=-2,037,<br>Effort: p=0,023, Z = -2.289 |
| IVIS+ADAS | *Dynamic Environment*<br>Performance: p=0,048, Z = -1,976 |

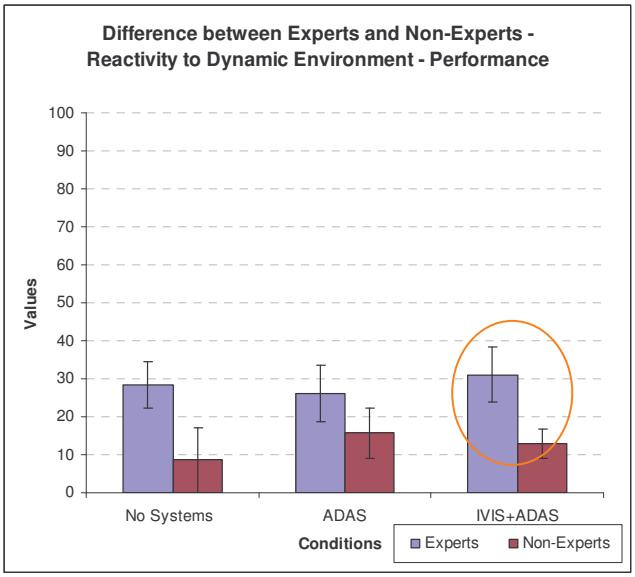



**Fig. 5-14: PSA-TLX - Differences between VR Experts and VR Non-Experts**

VR Non-Experts reported a lower level of Effort and disruption when driving in No System Condition than VR Experts in subtask *Reactivity to Static Environment*. VR Non-Experts also reported a lower effort when driving in IVIS+ADAS condition, as regards the *Reactivity to Dynamic Environment*.

### Overall Driving - Performance

All participants have been taken into account to evaluate Overall Driving Performance. Higher values represent a higher disruption, thus leading to a lower performance quality.

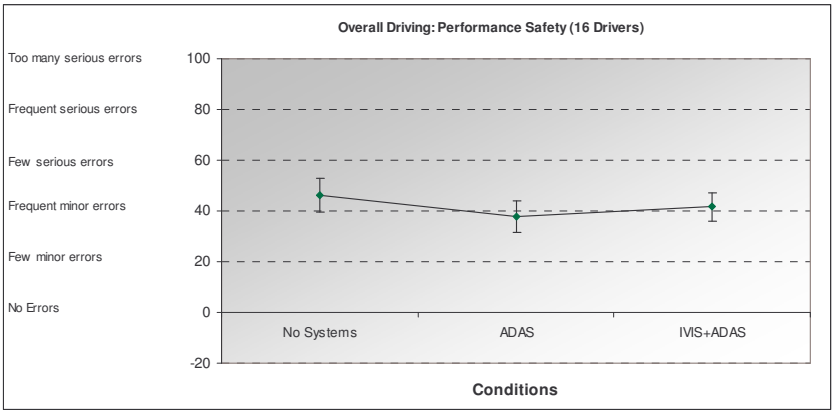



**Fig. 5-15: PSA-TLX - Overall Driving - Performance**

Wicoxon test showed no significant differences among Conditions.

### Overall Driving – Effort

All participants have been taken into account to evaluate Overall Driving Performance. Higher values represent a higher effort.

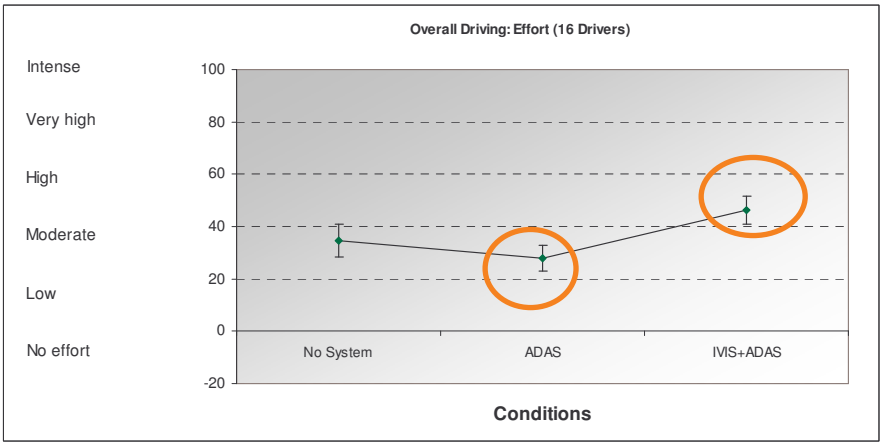



**Fig. 5-16: PSA-TLX - Overall Driving – Effort**

ADAS Condition and IVIS+ADAS Condition are statistically different. Effort in IVIS+ADAS Condition is reported as the heaviest, whilst the ADAS condition is the less demanding in terms of Effort ($p<0.011$).
No differences are reported between No System Condition and IVIS-ADAS Condition.

## Performance: Single Sub-Task score

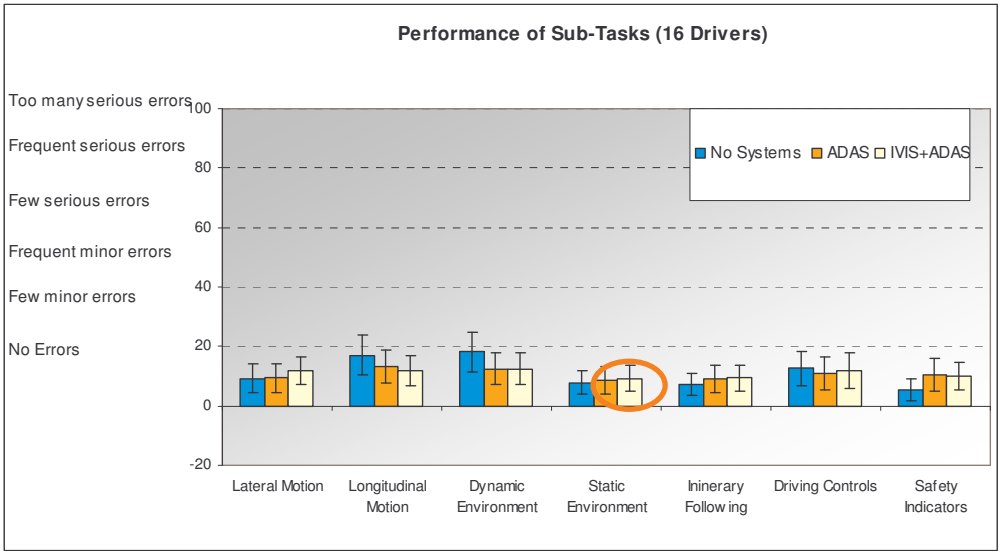



**Fig. 5-17: PSA-TLX – Single Sub-Tasks Score – Performance**

| Wilcoxon Test – Performance | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Confrontations** | **Lateral** | **Longitudinal** | **Dynamic** | **Static** | **Itinerary** | **Driving Commands** | **Safety Indicators** |
| No Systems vs. ADAS | 21,5 (NS) | 44 (NS) | 16 (NS) | 17,5 (NS) | 8,5 (NS) | 13 (NS) | 31,5 (NS) |
| ADAS vs. IVIS+ADAS | 22 (NS) | 31 (NS) | 36 (NS) | 23,5 (NS) | 20 (p<0,025) | 30 (NS) | 15,5 (NS) |
| No System vs. IVIS+ADAS | 56 (NS) | 36 (NS) | 39,5 (NS) | 48,5 (NS) | 19 (NS) | 37 (NS) | 30 (NS) |

**Table 5-3: PSA-TLX – Significances among pair comparison of Conditions for Performance**

A significant difference can be reported only between ADAS and IVIS+ADAS conditions, when dealing with Demand in Itinerary sub-task. The values of disruption are higher in the latter condition, meaning that participants saw their performance more degraded when they had to interact with the IVIS.

Detailed information about Performance distribution of scores for each single sub-task can be found in independent Appendix I.2.

## Performance management

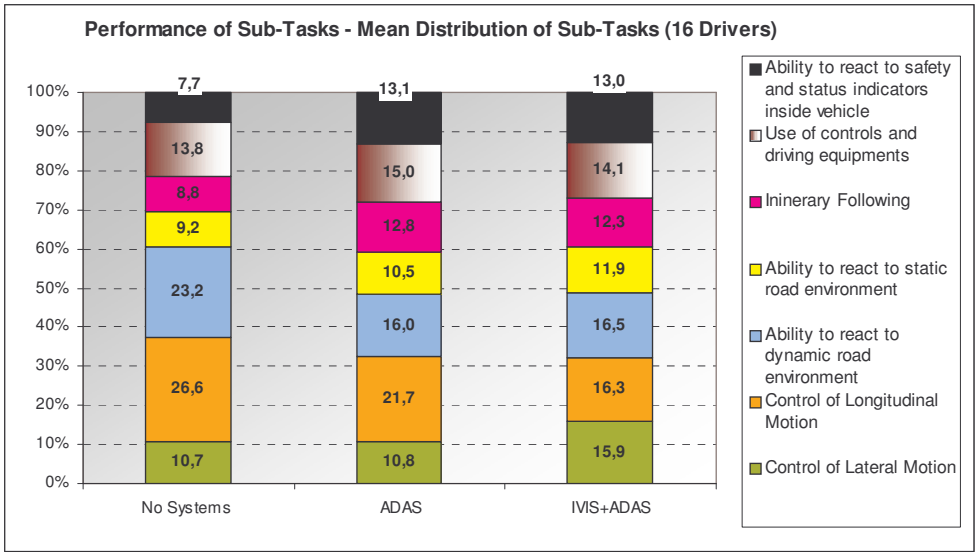



**Fig. 5-18: PSA-TLX – Performance components**

## Effort: Single Sub-Task score

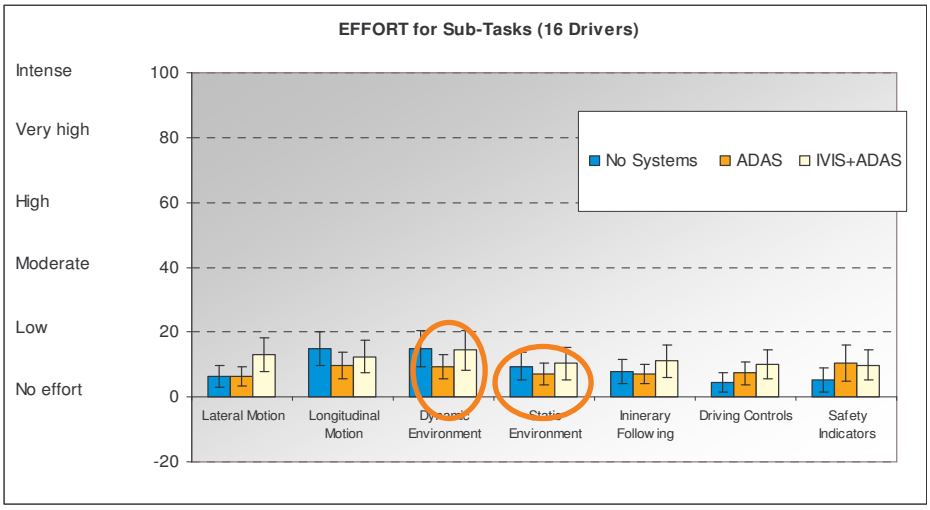



**Fig. 5-19: PSA-TLX – Single Sub-Tasks Score – Effort**

| Wilcoxon Test - EFFORT | | | | | | | |
|---|---|---|---|---|---|---|---|
| Confrontations | Lateral | Longitudinal | Dynamic | Static | Itinerary | Driving Commands | Safety Indicators |
| No Systems vs. ADAS | **11,5 (NS)** | **23 (NS)** | **16,5 (p<0,025)** | **8 (p<0,025)** | **23 (NS)** | **22,5 (NS)** | **9,5 (NS)** |
| ADAS vs. IVIS+ADAS | **18,5 (NS)** | **37 (NS)** | **50 (NS)** | **19 (NS)** | **44,5 (NS)** | **39,5 (NS)** | **32 (NS)** |
| No Systems vs. IVIS+ADAS | **45 (NS)** | **44,5 (NS)** | **43,5 (NS)** | **41,5 (NS)** | **30 (NS)** | **35 (NS)** | **22 (NS)** |

**Table 5-4: PSA-TLX – Significances among pair comparison of Conditions for Effort**

In both Reaction to Dynamic and Static Environment, participants were helped by ADAS to keep effort lower while driving in ADAS Condition than while driving in No System Condition (where no aid from ADAS was provided).
Detailed information about Effort distribution of scores for each single sub-task can be found in independent appendix I.2.

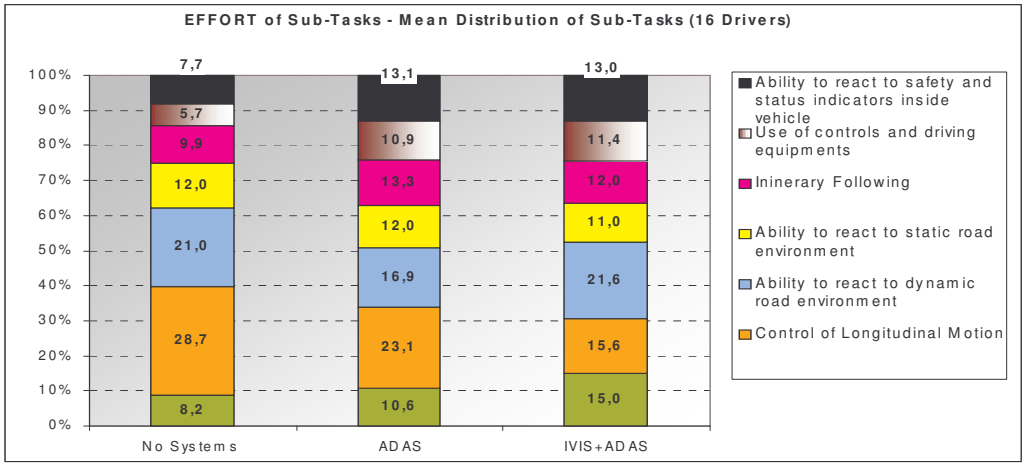



**Fig. 5-20: PSA-TLX – Mean distribution of Effort over Subtasks**

## Compromise Effort – Performance: Summary

Following the guidelines provided by PSA, Table 5-5 and Fig. 5-21 show a summarized representation of different Compromises for the three different experimental conditions.
The table represents how IVIS+ADAS and ADAS allow users to have a good or bad compromise.

| Compromise Level | No System vs. IVIS+ADAS | ADAS vs. IVIS+ADAS | No System vs. ADAS |
|---|---|---|---|
| Optimal | 5% | 5% | 7% |
| Very Favourable | 40% | 38% | 32% |
| Favourable | 18% | 23% | 38% |
| Unfavourable | 9% | 6% | 3% |
| Very Unfavourable | 28% | 28% | 21% |
| | 100% | 100% | 100% |

**Table 5-5: PSA-TLX – Compromise Effort-Performance; Summary of Distributions**

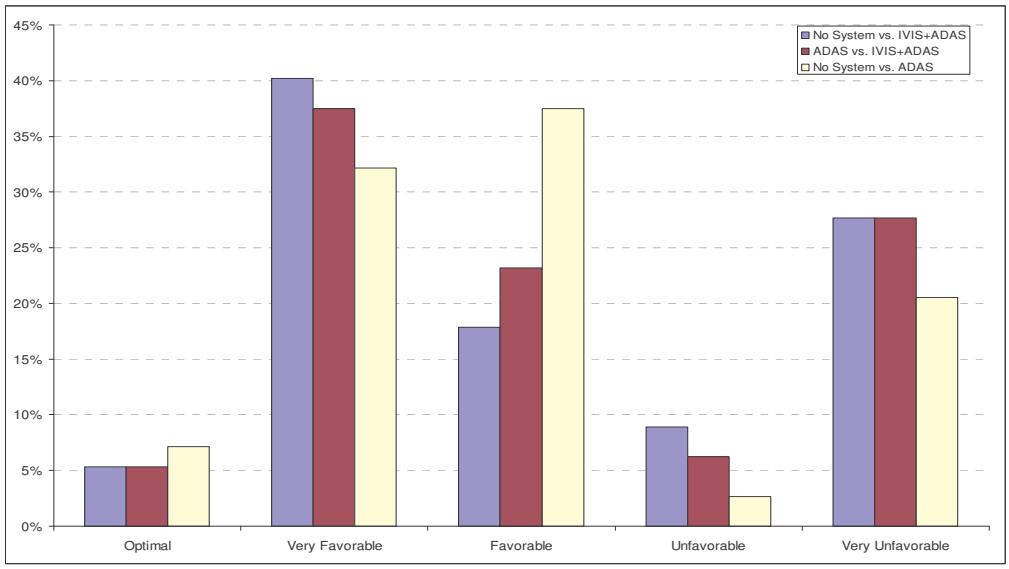



**Fig. 5-21: PSA-TLX – Graphical representation of Compromise levels**

The compromise shows how Effort and Performance vary when analyzing pairs of conditions. For more information, see independent appendix III.3.

Table 5-6 shows Synthetical Results of Compromises Levels in percentage for each pair of Conditions. In the main part of the cases, Very Favourable Compromise is attained. This permits to say that the ADAS allows to keep the effort and the disruption while driving at low values, thus raising either Performance and safety during driving. Moreover, the Very Unfavourable Compromise is attained for almost 1 out of 4 participants when comparing to IVIS+ADAS Condition, and for almost 1 out of 5 participants when comparing No System to ADAS.

| | Lateral Control | | | Longitudinal Control | | | Dynamic Environment | | | Static Environment | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | NS vs. A | NS vs. IA | A vs. IA | NS vs. A | NS vs. IA | A vs. IA | NS vs. A | NS vs. IA | A vs. IA | NS vs. A | NS vs. IA | A vs. IA |
| Good | 8 | 10 | 12 | 12 | 11 | 12 | 9 | 9 | 12 | 11 | 12 | 13 |
| Bad | 8 | 6 | 4 | 4 | 5 | 4 | 7 | 7 | 4 | 5 | 4 | 3 |

| | Itinerary Following | | | Driving Commands | | | Safety Devices | | |
|---|---|---|---|---|---|---|---|---|---|
| | NS vs. A | NS vs. IA | A vs. IA | NS vs. A | NS vs. IA | A vs. IA | NS vs. A | NS vs. IA | A vs. IA |
| Good | 11 | 12 | 13 | 9 | 10 | 13 | 11 | 10 | 11 |
| Bad | 5 | 4 | 3 | 7 | 6 | 3 | 5 | 6 | 5 |

**Table 5-6: PSA-TLX – Rating of Compromise Quality among Sub-tasks**

**Good** includes Optimal, Very Favourable and Favourable compromises;
**Bad** includes Unfavourable and Very Unfavourable compromises.

IVIS+ADAS condition always presents a significant Good level of compromise, whereas ADAS condition presents a significant Good level of compromise in Longitudinal Control, Static Environment and Driving Commands contexts.

### 5.2.2 DALI results

Statistical correlation tests have been calculated between raw values and weighted values.

**DALI Factors across experimental conditions – VR Experts vs. VR non-Experts**

No significant differences between VR Experts and VR non-Experts in No System Condition have been detected, except for Stress level, where just a tendency can be reported.
No significant differences between VR Experts and VR non-Experts in ADAS Condition and in IVIS+ADAS Condition have been detected.

**Correlation between Weighted and Raw Values**

The Test of Correlation of Pearson shows a very high relationship between Weighted and Raw values.

| Weighted vs. Raw Correlation | Pearson Correlation Coefficient for Factor… | | | | | |
|---|---|---|---|---|---|---|
| | Attention Demand | Visual Demand | Auditive Demand | Temporal Demand | Interference | Stress |
| | No System Condition | | | | | |
| **Correlation** | ,900(**) | ,950(**) | ,923(**) | ,587(*) | ,596(*) | ,966(**) |
| **Sig. (2-queue)** | 0,000 | 0,000 | 0,000 | 0,017 | 0,015 | 0,000 |
| | ADAS Condition | | | | | |
| **Correlation** | ,971(**) | ,845(**) | ,728(**) | ,898(**) | ,915(**) | ,807(**) |
| **Sig. (2- queue)** | 0,000 | 0,000 | 0,001 | 0,000 | 0,000 | 0,000 |
| | IVIS+ADAS Condition | | | | | |
| **Correlation** | ,849(**) | ,760(**) | ,807(**) | ,897(**) | ,728(**) | ,848(**) |
| **Sig. (2- queue)** | 0,000 | 0,001 | 0,000 | 0,000 | 0,001 | 0,000 |

**Table 5-7: DALI - Relationship between Weighted and Raw Data**

Correlations in every Condition are very high, and it can be stated that the same results and meanings can be achieved by using either Raw and Weighted data in this experiment.

**DALI Global and Factors Score – Weighted Data**

In Fig. 5-22, that represents the DALI Factors for all the sample, all participants have been taken into account, as no significant difference between Experts and non-Experts has been found.

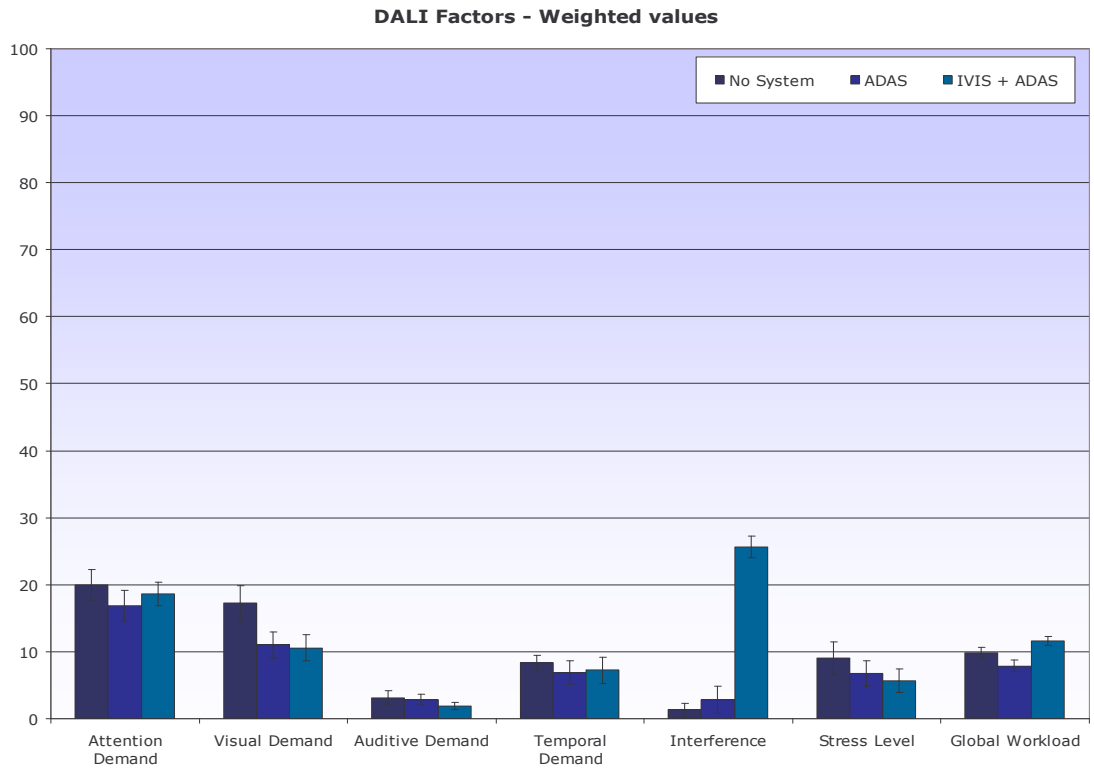



**Fig. 5-22: DALI - Difference among Conditions respect to different Factors and Global Workload and summarizing table with Wilcoxon Text**

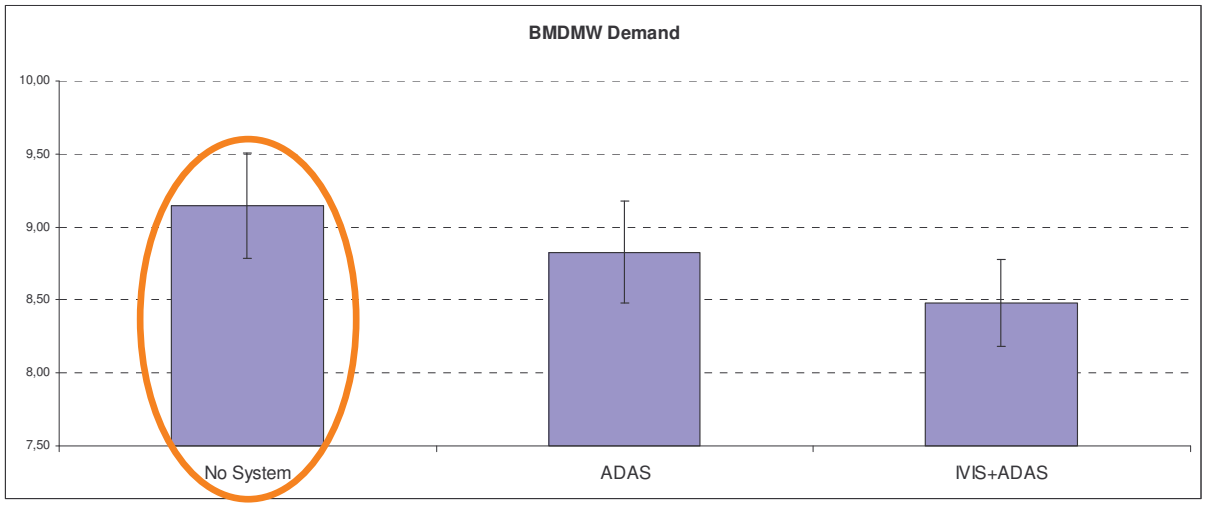
| Global Demand | | | Stress | Interference | | Visual Demand | |
|---|---|---|---|---|---|---|---|
| NS – A | NS - IA | A - IA | NS - IA | NS-IA | A-IA | NS-A | NS-IA |
| -2,442 | -2,224 | -3,155 | -2,197 | 3,544 | -3,43001 | -2,733 | -2,133 |
| p=0,015 | p=0,026 | p=0,002 | p=0,028 | p=0,000 | p=0,0006 | p=0,006 | p=0,033 |

**Table 5-8: DALI – Significances among DALI factors for different conditions**

Significant differences across the three conditions in Global Demand have been detected, and can be analytically seen in the table above. According to participants, the maximum load was experienced when interacting with the IVIS, whereas the ADAS Condition resulted in being the least demanding. The No System Condition stands between the other conditions.
Visual Demand has its peak in No System Condition, significantly different from the other conditions, as the aid of ADAS is not provided. No difference between ADAS and IVIS+ADAS condition are detected, meaning that the use of IVIS in the IVIS+ADAS condition is not sufficient to conclude it gives significant variation from ADAS condition.
Interference is significantly higher in IVIS+ADAS condition rather than the other conditions, because the major interference from the IVIS was present only in this Condition. In fact, no significant difference was detected between No System and ADAS conditions.
Stress is felt significantly higher when driving in No System Condition rather than in ADAS condition. The ease given by the ADAS is missing in No System condition, and this miss has direct reflection on perceived stress.

### 5.2.3 BMDMW results

The value of each Factor is computed as the sum of values of items composing that Factor.
Global Demand of Conditions is computed as the mean of values of single Factors.
Concerning Global Demand in Fig. 5-23, higher values in the graphics indicate a higher load level, therefore a worse driving performance.




| | No System - ADAS | No System - IVIS+ADAS | ADAS - IVIS+ADAS |
|---|---|---|---|
| Z | -2,01901 | -2,82734 | -1,76165 |
| Sig. Asint. a 2 code | p=0,043487 | p=0,004694 | p=0,078128 |

a Based upon positive ranks.

**Fig. 5-23: BMDMW – Global Demand across Conditions**

**Discussion**

From Wilcoxon Test, it appears that ADAS Condition and IVIS+ADAS Condition are significantly less demanding than No System Condition. The latter statement may be explained by the use of ADAS, which mitigates the negative aspect of perceived stress from the interaction with the IVIS. No significant differences were found between ADAS and IVIS+ADAS Condition.

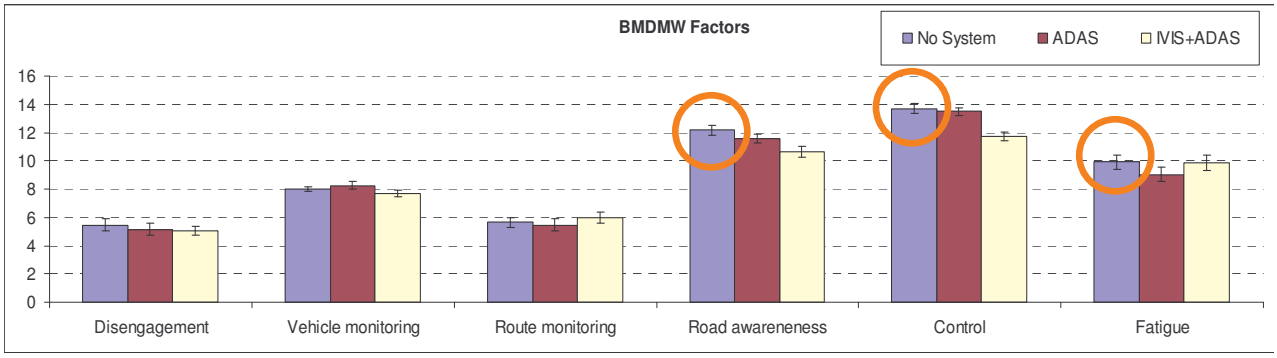



**Fig. 5-24: BMDMW – Factors of Demand across Conditions**

As regards the analysis of single Factors, only the following differences were found:

| | Wilcoxon Test | | | | |
|---|---|---|---|---|---|
| | Road Awareness | | Control | | Fatigue |
| | No System - IVIS+ADAS | ADAS - IVIS+ADAS | No System - IVIS+ADAS | ADAS - IVIS+ADAS | No System - ADAS |
| Z | -2,862 | -2,407 | -3,018 | -3,106 | -2,286 |
| Sig. Asint. a 2 code | p=0,004 | p=0,016 | p=0,003 | p=0,002 | p=0,022 |

Based upon positive ranks.

**Table 5-9: BMDMW – Significances among BMDMW factors for different conditions**

**Discussion**
It is here important to remind that higher values of Factors mean a worse driving performance.
Participants report Road Awareness in IVIS+ADAS Conditions as the highest. This can be explained by the fact that participants interacting with the IVIS kept a lower speed and watched more times at the road, thus having a higher Road Awareness.
Participants also report a higher control in IVIS+ADAS Condition, perhaps because of slower speed and higher Road Awareness.
Last, Fatigue is reported by participants as lower in ADAS Condition rather than in No System Condition.

### 5.2.4 Objective results

Data collected via the VR Simulator was analyzed as described in § 5.1.5. Results are shown in the remaining of this paragraph.

Raw Data graphics are used to show variability of Normalized Indices components. This is legitimate as the two distributions are the same, so it is not necessary to operate a transformation for all distributions, as far as the variability can be deduced by its graphical representation.

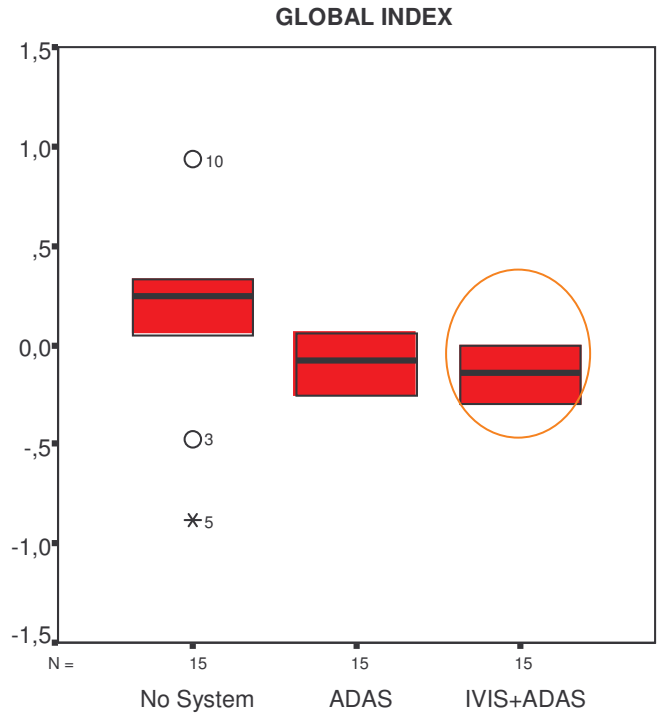
The red box indicates the variability of the sample in the different Conditions, the horizontal black line indicates the mean value. Out layers are marked with an O or an asterisk, indicating participants who displaced of more than one Standard Deviation from the mean.
As described in § 5.1.5, lower values are better because closer to the baseline.

After having verified homoschedasticy of variances and the Normal distribution of data, a parametric test was applied: univariate ANOVA. Moreover, to carry out the pair-comparison between conditions, Bonferroni post-hoc test was used.

The Global Index across conditions is presented first. Then, Synthetical Indices for each segment are presented and described. At the end a deep analysis of secondary task demand will be reported in order to better understand indexes results.

Tables with statistical differences among conditions are provided. Where pairs of conditions are missing, no significant statistically difference has been found. When just one significant difference is available, it is included in the graphic.

**Global Variability Index**




| Bonferroni Test: Significance of Differences | |
|---|---|
| **No System - ADAS** | **No System - IVIS+ADAS** |
| p=0,057 (Tendency) | p=0,027 |

**Fig. 5-25: Global Index of Conditions**

The IVIS+ADAS Condition appears to be the least variable, whereas the No System Condition presents the highest variability and displacement from baseline. The ADAS Condition is not significantly different from the IVIS+ADAS condition, whilst statistical differences are present between the other pairs, as described in Fig. 5-25.

**Discussion**

Two considerations can be done to explain the trend depicted above:

- The use of ADAS allows subjects to decrease variability of different metrics during driving task;
- The awareness of having to execute a difficult task in IVIS+ADAS Condition puts users in a high load state. This state has the effect to lower variability of different metrics during driving task.

**Brake Segment – Synthetic Variability Index**

The metrics which compose the Synthetical Variability Index for Brake Segment are the following:

- Warning duration (red/yellow)
- Brake reaction time
- Action on brake

The metrics were computed for each Visibility condition (150, 60 and 30 metres respectively).

| Significance of Difference (Bonferroni Test) | NS - ADAS | ADAS – IVIS+ADAS | NS - IVIS+ADAS |
|---|---|---|---|
| **Good Visibility** | 0.001 | 0.001 | - |
| **Medium Visibility** | 0.003 | - | 0.002 |
| **Bad Visibility** | - | 0.038 | |

**Table 5-10: Objective Results – significances in Brake Segment**

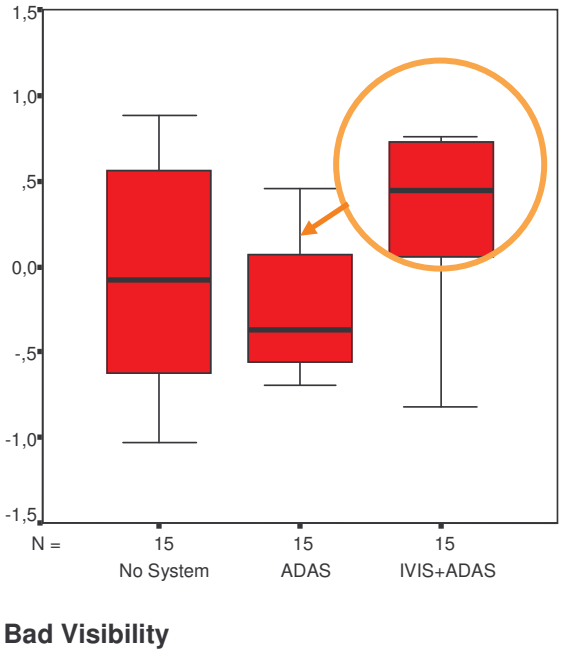
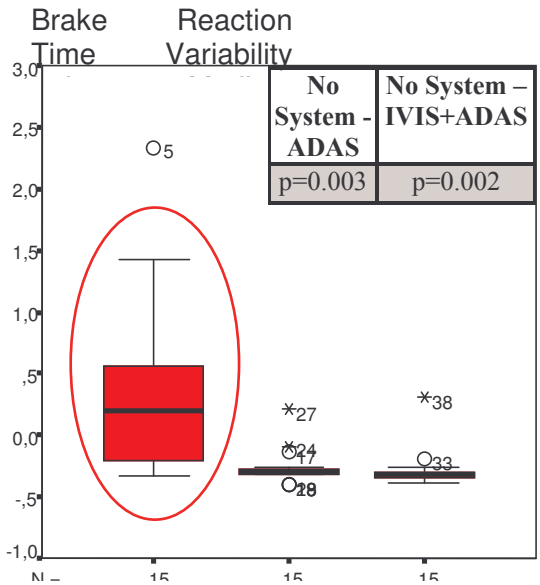
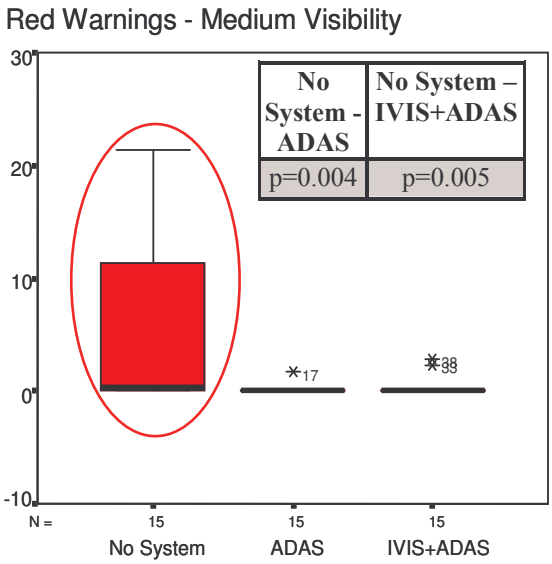



**Fig. 5-26: Objective Results - Synthetical Variability Indices for Brake segment**

As shown in Table 5-10, in Bad visibility the IVIS+ADAS condition is different from ADAS condition, indicating that the Braking actions were more difficult in the first condition. Medium and Good Visibility are further explained by Warning metrics, with Bonferroni statistical method to show differences among conditions as regards Warning metrics.



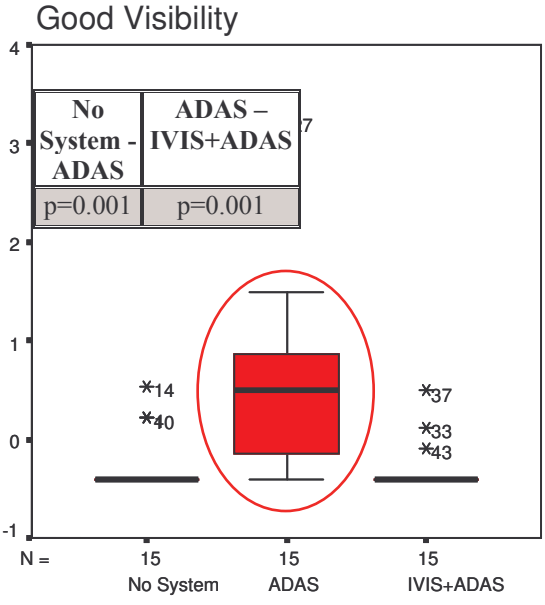

Brake Reaction Time Variability Index in Medium Visiblity (a)



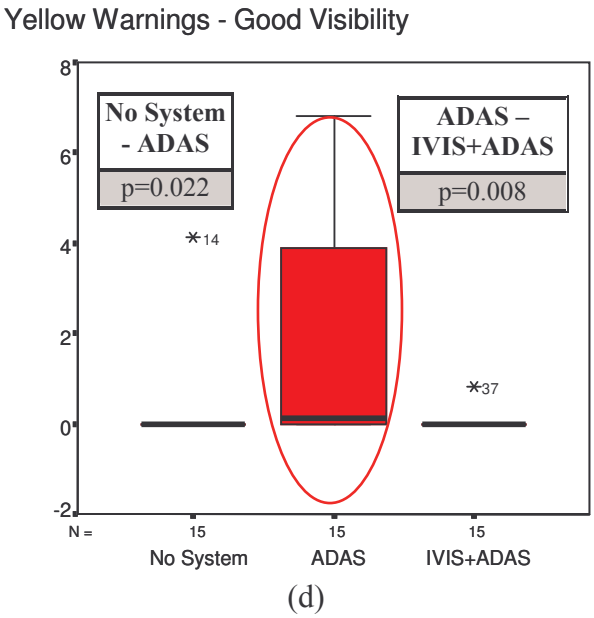

Variability of Red Warning in Medium Visibility (b)

**Fig. 5-27: Objective Results – Variability of Brake reaction time and Warning duration in Brake segment – Medium Visibility**

In Medium Visibility, as shown in Table 5-10, No System Condition presents a higher variability (a) in actions related to braking respect to the other two conditions. This difference can be explained by looking at figure (b): people tend to stay significantly longer in a Red Warning state, during No System condition, so limited visibility conditions lead drivers to get closer to leading vehicles.




(c)




(d)

**Fig. 5-28: Objective Results - Variability of Brake reaction times and Warning duration in Brake segment – Good Visibility**

In Good Visibility, as show in Table 5-10, ADAS Condition presents a higher variability (c) in actions related to braking respect to the other two conditions. This difference can be explained by looking at figure (d): people tend to stay significantly longer in a Yellow Warning state, during ADAS condition, leading drivers to get closer to leading vehicles.

**Discussion**

ADAS seems to help participants to better keep a safety distance from leading vehicles. In Good Visibility conditions, permanence in Yellow Warning state can be explained as a trust in the system, as other metrics are not significantly different among the conditions (such as Brake reaction time and Action on brake).

When no ADAS is available to the driver, as in the case of No System Condition, participants are more likely to enter the Red Warning state, thus getting in a potentially dangerous situation as regards their speed and relative distance from leading vehicles.

### Shrinkage Segment – Synthetic Variability Index

The metrics which compose the Synthetical Variability Index for Shrinkage segment are the following:

- Warning duration (red/yellow)
- Brake reaction time
- Action on brake
- Headway
- Relative Distance

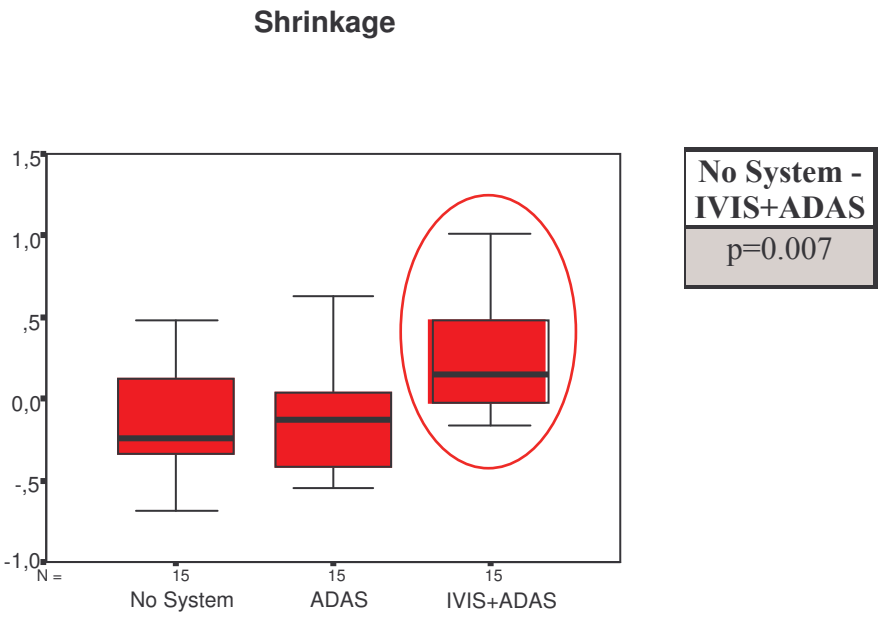



**Fig. 5-29: Objective Results - Synthetical Variability Indices for Shrinkage segment**

Inside the Shrinkage segment, IVIS+ADAS Condition appears to be significantly different from No System Condition, as can be seen in Fig. 5-29. No difference can be found among the other conditions. Detailed information of the composition of this variability can be found in the following graphics:

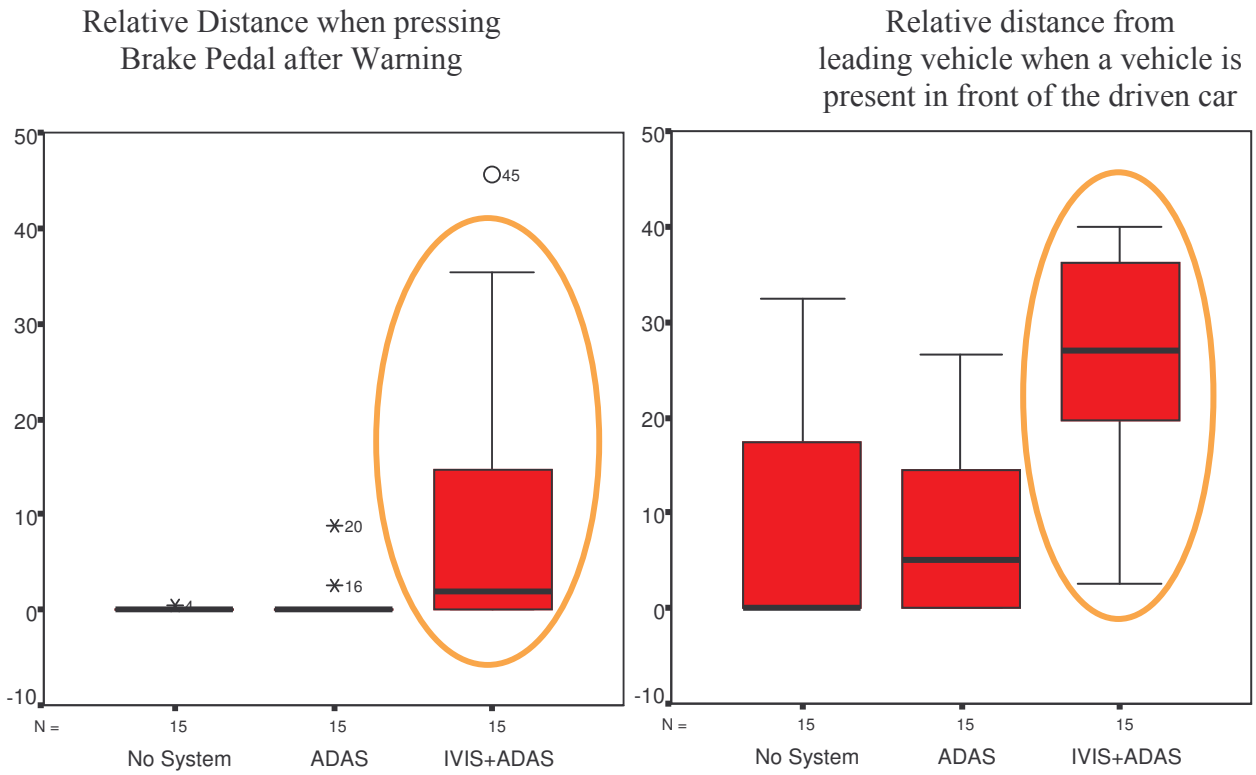



**Fig. 5-30: Objective Results – Relative distances variability for Shrinkage segment**

| Significance of Difference (Bonferroni Test) – Relative Distance… | … when pressing Brake pedal after Warning | | … from leading vehicle (when any) | |
|---|---|---|---|---|
| | No System – IVIS+ADAS | ADAS – IVIS+ADAS | No System – ADAS | ADAS – IVIS+ADAS |
| | p=0.015 | p=0.032 | p=0.000 | p=0.000 |

**Table 5-11: Objective Results - Analytical Significances in Shrinkage Segment**

For IVIS+ADAS, the Relative Distance Variability when pressing Brake Pedal after Warning results significantly higher respect the other two conditions. On the other hand, relative distance variability from a leading vehicle is significantly higher in IVIS+ADAS conditions from the one recorded for ADAS condition. No difference is reported between No System and IVIS+ADAS Condition.

**Discussion**

It can be concluded that the use of an IVIS has a negative impact on reactivity to ADAS warnings, which could be translated into a potentially dangerous situation during driving task. This statement can be justified as follows: after being warned by the ADAS, participants reacted less regularly on brake pedal according to the leading vehicle, thus having high variability in the "Relative distance when pressing Brake Pedal after Warning" metric. As a confirmation of results related to Relative distance when pressing Brake Pedal after Warning, also the "Relative distance from leading vehicle" metric shows a significant difference

between IVIS+ADAS condition and the other conditions. This can be explained by the interaction with the IVIS, raising reaction time when the warning pops up.

**Obstacle Segment – Synthetic Variability Index**

The metrics which compose the Synthetical Variability Index for Obstacle segment are the following:

- Warning duration (red/yellow)
- Brake reaction time
- Action on brake

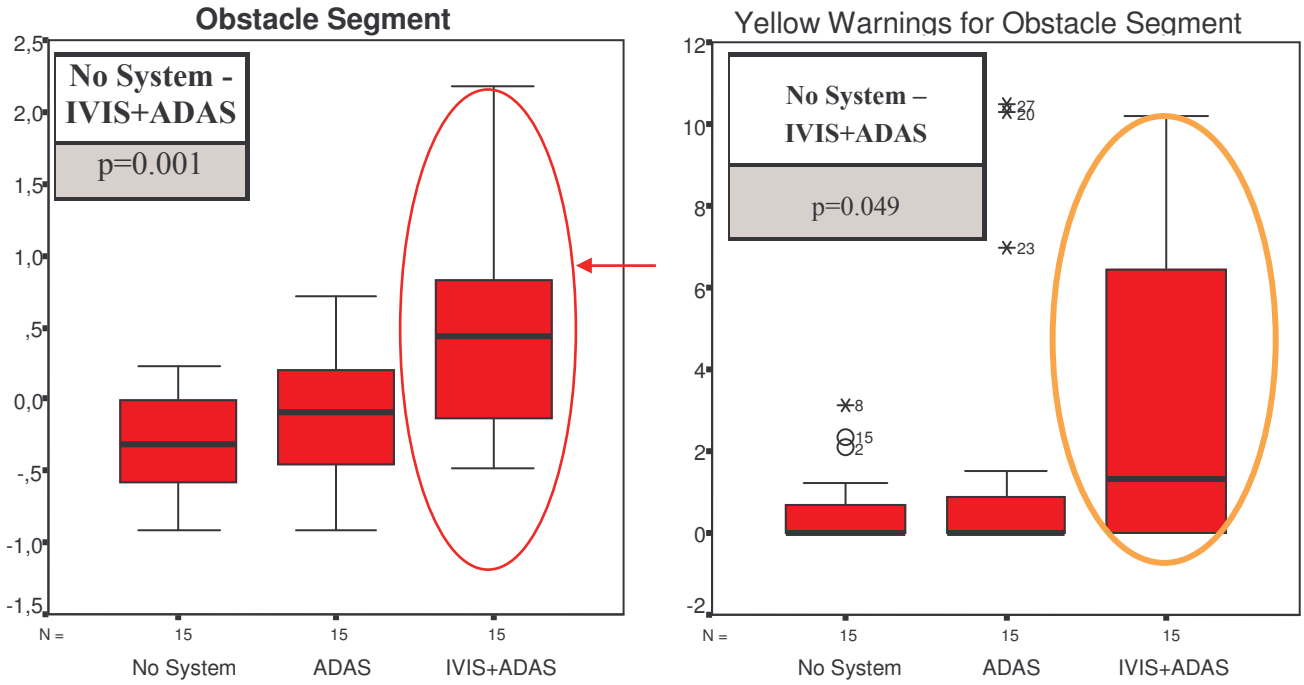



**Fig. 5-31: Objective Results – Synthetical Variability Indices for Obstacle segment**

The IVIS+ADAS condition appears to be significantly different from the No System condition when encountering the obstacle on the lane. Participants realize later that there is a stationary vehicle on the lane, thus braking later. In fact, permanence in Yellow ADAS warnings is significantly higher in IVIS+ADAS condition. This effect derives from Demand effect due to task execution.
In other conditions (No System and ADAS Conditions), people began the maneuver of avoidance earlier, thus reducing time in Yellow Warning areas and seldom entering Red Warning areas.

No significant differences have been found between the other pairs of conditions.

**Discussion**
The effect of the IVIS during the drive has a negative impact on reaction times to ADAS Warnings, thus leading to potentially dangerous situations during the drive, as can be seen by a higher permanence in the Yellow zone before realizing that the vehicle in front is stationary, and an avoiding maneuver is required.

**Unexpected Action Segment – Synthetic Variability Index**

The metrics which compose the Synthetical Variability Index for Unexpected Action segment are the following:

- Warning duration (red/yellow)
- Brake reaction time
- Action on brake

No statistically significant differences were detected in this segment.

**Deeper analysis of secondary task impact**

The metrics used to study Lateral Behaviour in Task Segment for the three experimental conditions are the following:

- Mean Speed
- Lateral Displacement

The metrics were computed for each Visibility condition (150, 60 and 30 meters respectively).

No index was computed for Lateral Behaviour, as simple study of Mean Speed SD and Lateral Displacement SD was sufficient to highlight significant differences across experimental conditions and Visibility levels.

As the distribution of data related to the metrics in the Task segment was not homoschedastic, to carry out the post-hoc statistical analysis the Games-Howell statistical method was used instead of Bonferroni.

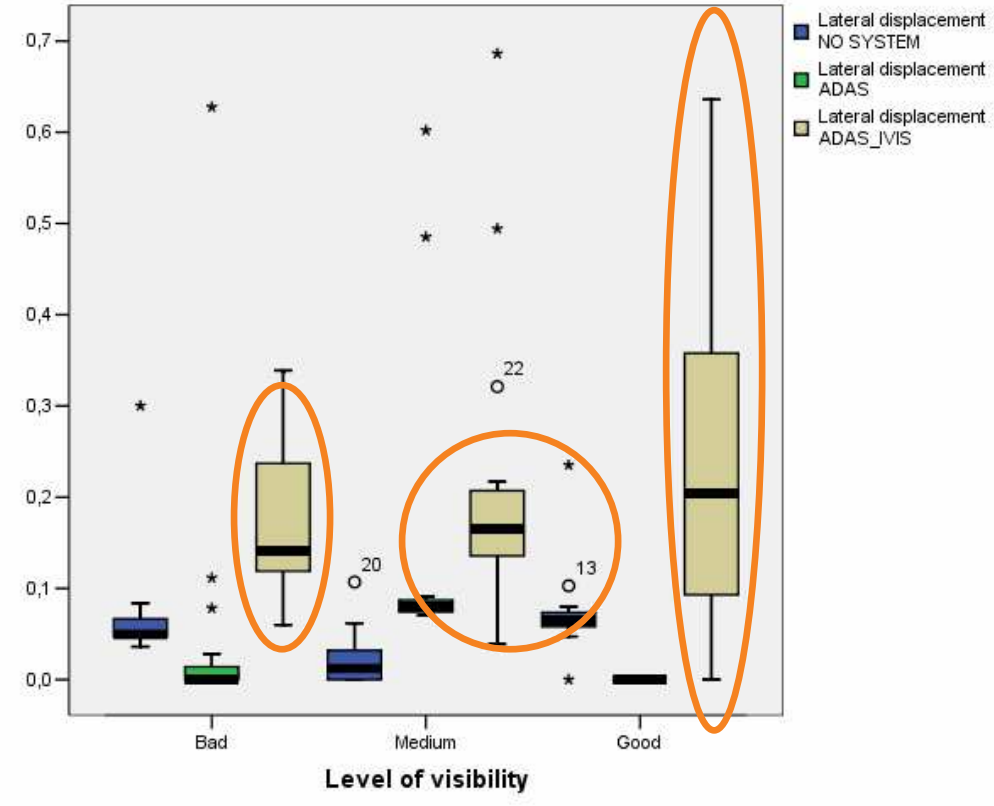


**Fig. 5-32: Deeper analysis of secondary task impact – Lateral displacement variations**

| Lateral Displacement | No System - ADAS | No System – IVIS+ADAS | ADAS – IVIS+ADAS |
|---|---|---|---|
| Good Visibility | 0,000 | | |
| Good Visibility | | 0,022 | |
| Good Visibility | | | 0,001 |
| Medium Visibility | 0,039 | | |
| Medium Visibility | | 0,002 | |
| Bad Visibility | 0,004 | | 0,034 |

**Table 5-12: Objective Results – Lateral displacement significances in Task Segment**

Lateral displacement variability appears to be higher in IVIS+ADAS condition than the other two conditions; in particular, in Bad visibility it is significally higher than ADAS condition, in Medium visibility it is significally higher than No system condition and in Good visibility it is significally higher than No System and ADAS condition.
The No System condition variability appears to be higher than ADAS condition in Good and Bad visibility. The only case in which the ADAS condition variability is not the lowest is respect to the No System condition in Medium visibility.
There is no significant difference in ADAS and IVIS+ADAS condition in Medium visibility.

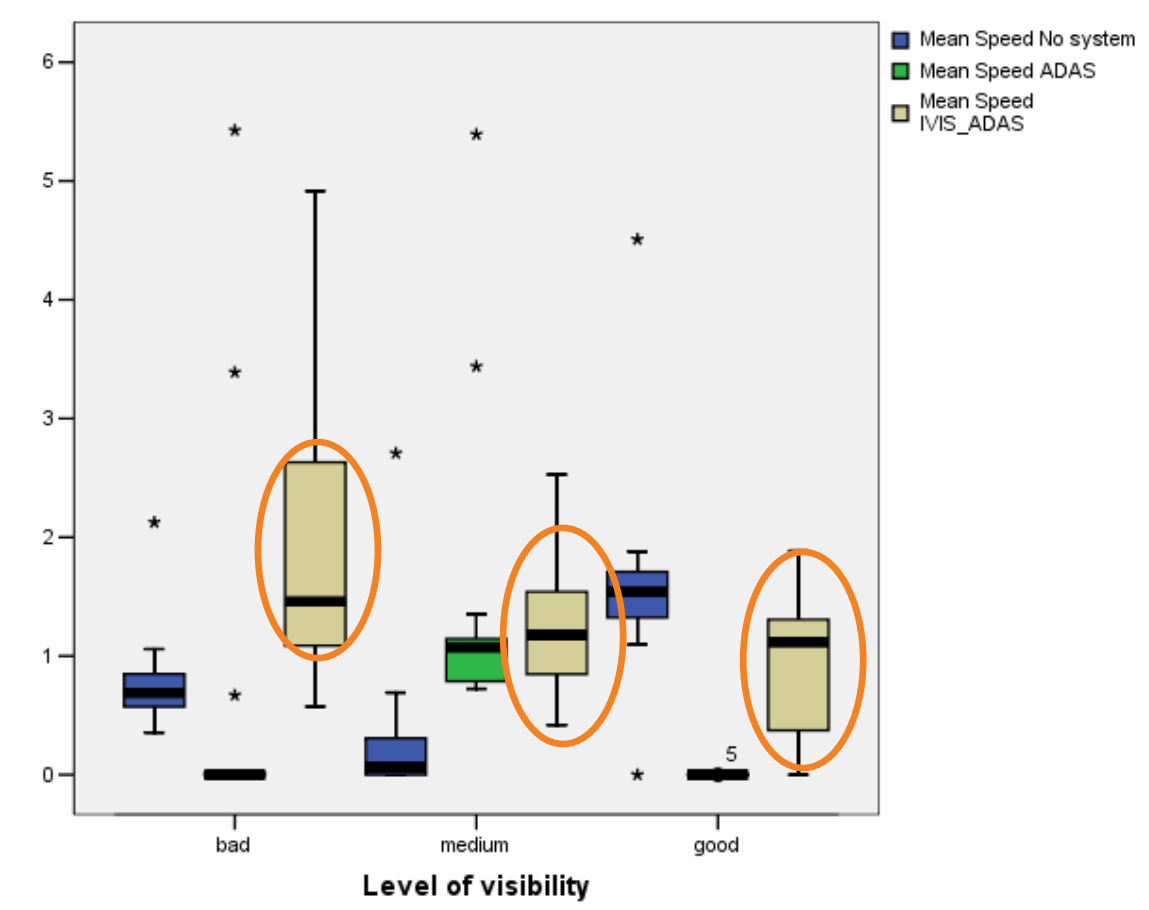



**Fig. 5-33: Deeper analysis of secondary task impact – Mean speed variations**

| Mean Speed | No System - ADAS | No System – IVIS+ADAS | ADAS – IVIS+ADAS |
|---|---|---|---|
| **Good Visibility** | | 0,056 (tendency) | 0,000 |
| **Medium Visibility** | 0,022 | 0,002 | |
| **Bad Visibility** | | 0,011 | 0,059 (tendency) |

**Table 5-13: Objective Results – Mean speed significances in Task Segment**

Mean speed variability for IVIS+ADAS condition appears to be significantly higher than No System condition in Medium visibility and Bad visibility, and it is significantly higher than ADAS condition in Good visibility. In Bad visibility, no significant difference was found between IVIS+ADAS condition and No System condition.
In Medium visibility, the ADAS condition appears to be significantly higher than the No System condition.

**Discussion**
In general, during the IVIS+ADAS condition, more variability in Lateral displacement and Mean speed is reported, due to the interaction with the IVIS, in all visibility states. It can be therefore stated that the interaction with the IVIS, even with the aid of the ADAS, has a negative impact on driving performance.

When ADAS condition presents a higher Lateral displacement respect to the No System condition while driving in medium visibility, a higher variability in Mean speed is also reported.

### 5.2.5 Comparison between Objective and Subjective results

Objective Indices have been compared with results from subjective tools, to verify whether questionnaires reported a trend similar to the objective data collected. Points of contact are reported in the remaining of this paragraph, divided by each subjective tool.

**PSA-TLX Comparison**

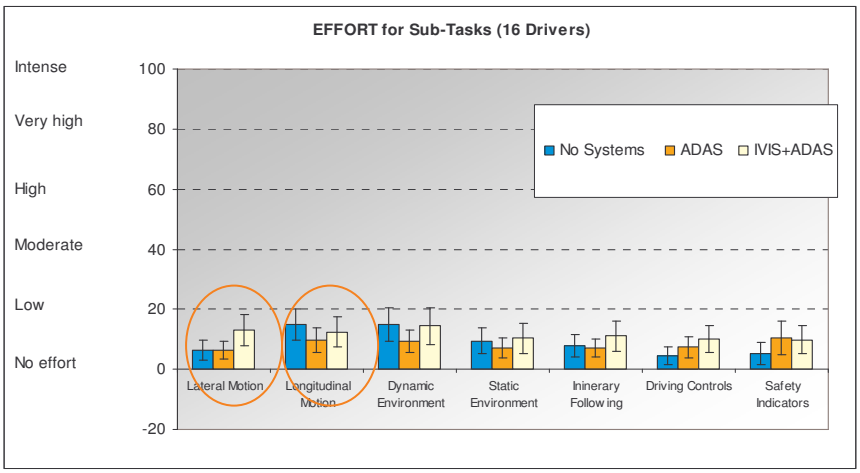



**Fig. 5-34: Comparison between PSA-TLX results and Objective Results - Effort**

As regards Effort in Reactivity to Dynamic and Static Environment, PSA-TLX reports a significant higher effort for No System Condition respect to ADAS condition. This trend is

the same, in terms of variability from the baseline, in the objective results between No System and ADAS condition.

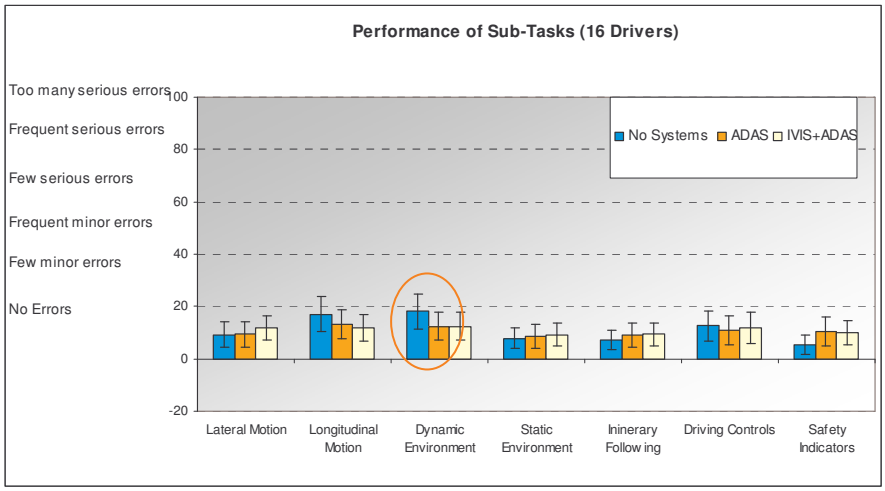



**Fig. 5-35: Comparison between PSA-TLX results and Objective Results - Effort**

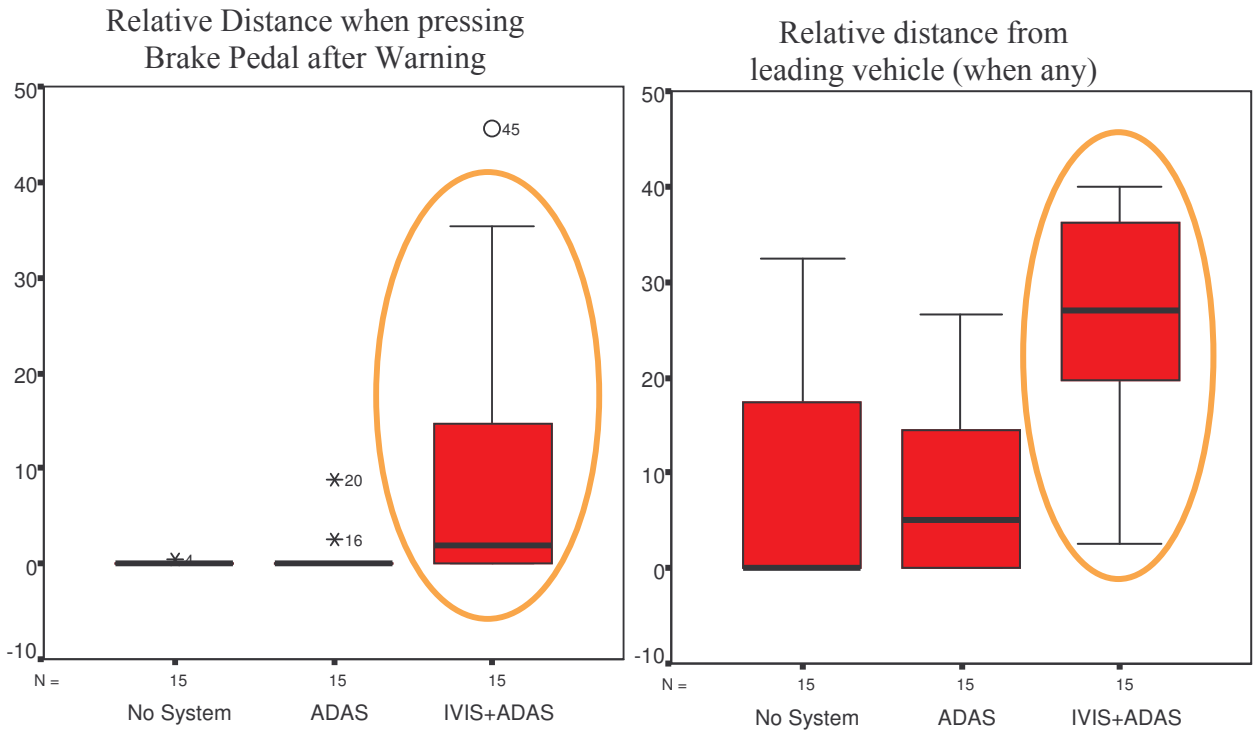



**Fig. 5-36: Comparison between PSA-TLX results and Objective Results - Performance**

In Itinerary Following sub-task, PSA-TLX reports a significant difference between ADAS and IVIS+ADAS conditions. This fact is confirmed by objective data analysis in Shrinkage segment, where the ADAS conditions results significantly less varying than the IVIS+ADAS condition as regards Relative distance when pressing Brake pedal after warning and Relative distance from leading vehicles.

## DALI Comparison

**Fig. 5-37: Comparison between DALI Visual Demand and Objective Results**

- Visual Demand as recorded with DALI appears higher in No System condition. This aspect is confirmed in objective results, where situations of longer permanence in ADAS Warnings sections (Brake segments), In such situations, the Visual load is higher as the particular situation requires more visual attention.

- DALI reports a significant higher level of Interference in IVIS+ADAS condition. Situations in which objective results present the same trend can be found in overall Brake segment in bad visibility, during Obstacle Segment an during Shrinkage Segment.

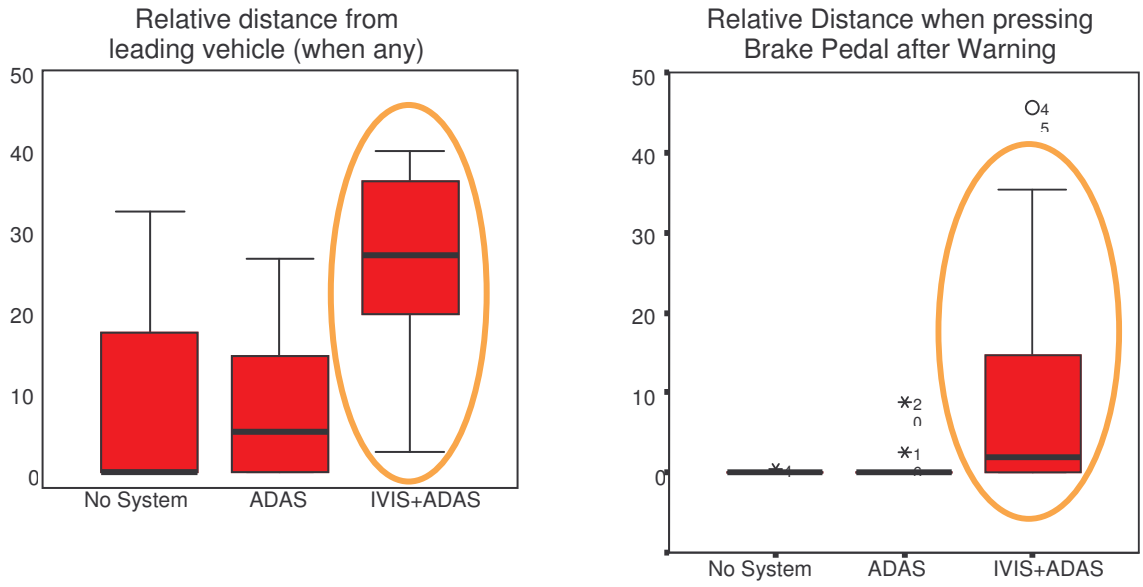



**Fig. 5-38: Comparison between DALI Interference and Objective Results**

- The DALI value of stress is significant higher in No System condition rather than IVIS+ADAS condition. Situations in which the same trend is reported in objective results are the Brake segment variability in Medium Visibility, and the Red Warnings Duration variability in Medium Visibility.

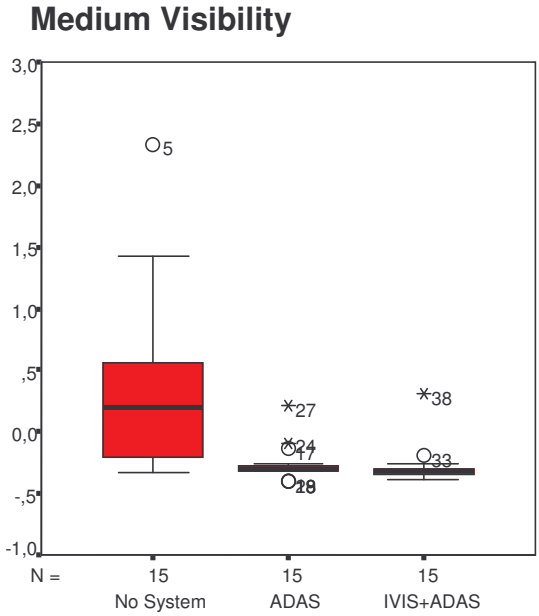
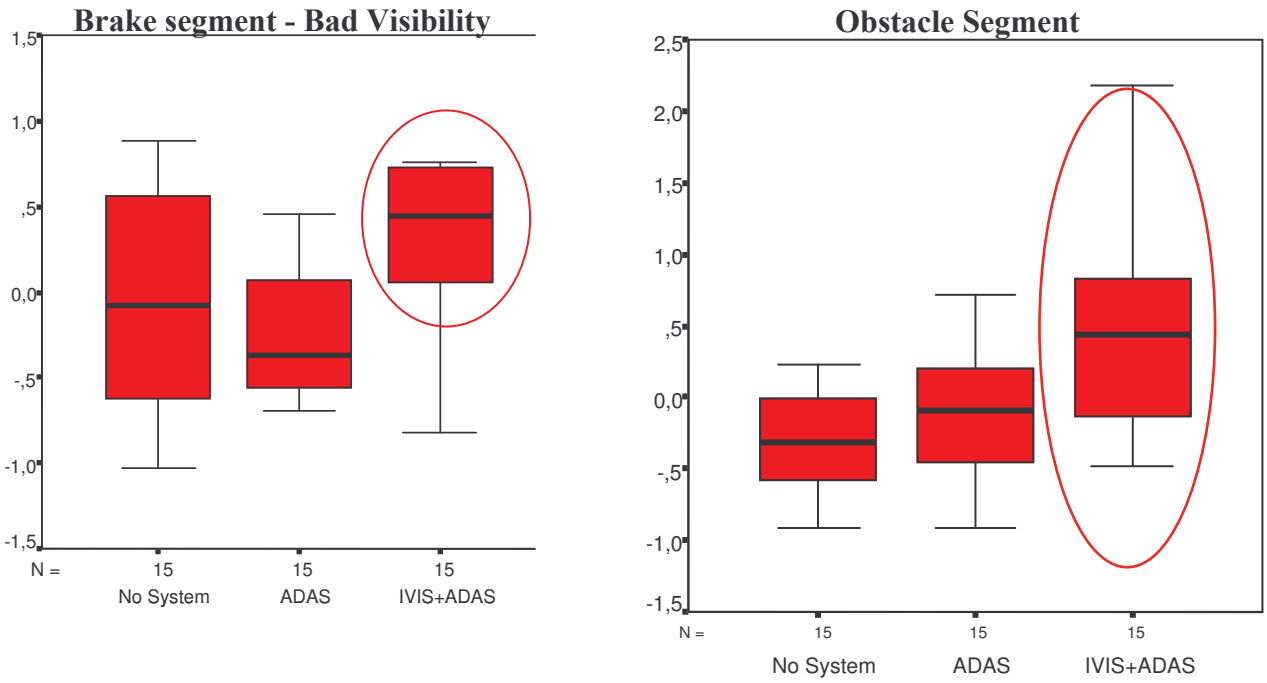



**Fig. 5-39: Comparison between DALI Stress and Objective Results**

- At global level, ADAS condition is the least demanding, and the No System Condition is less demanding than the IVIS+ADAS condition. This trend is confirmed by objective results in:
  - Brake segment, which shows a better performance in ADAS condition rather than No System condition in Medium Visibility;

- Shrinkage segment and Obstacle segment, No System condition shows lower variability than IVIS+ADAS condition



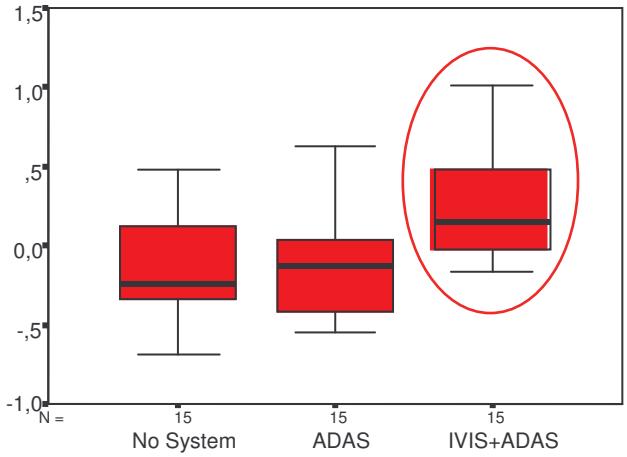
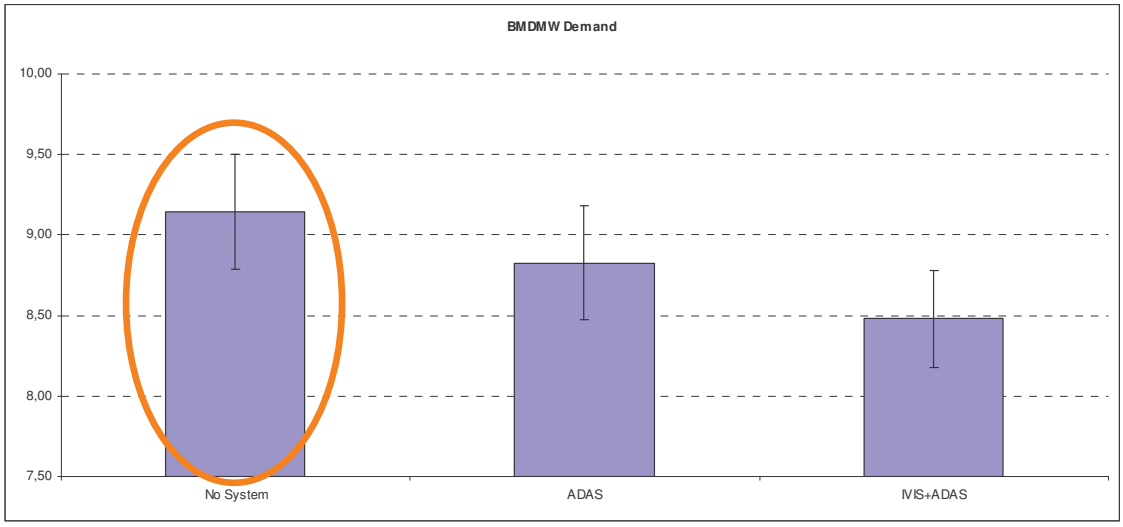
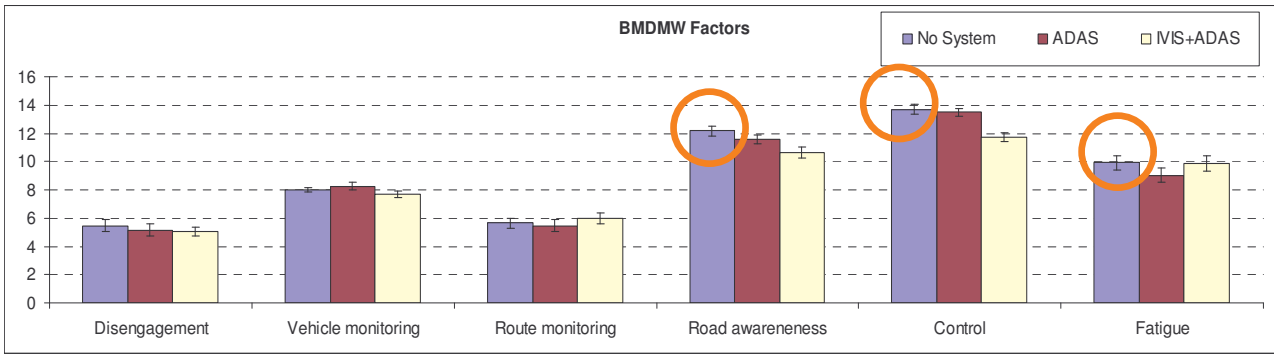




**Fig. 5-40: Comparison between DALI Global Demand and Objective Results**

## BMDMW Comparison



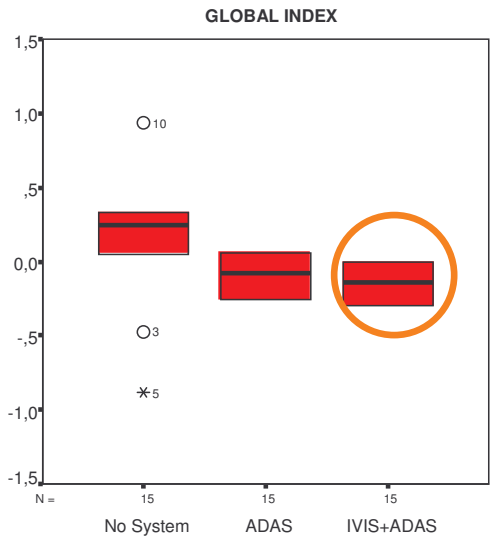




**Fig. 5-41: Comparison between BMDMW Results and Objective Results**

The trend reported by BMDMW is confirmed by Objective Data Analysis by the Global Variability Index. At Factors level, Road Awareness shows the same trend of Objective Data Analysis, because lower variations in Objective Data may correspond to a higher Road Awareness.

Control Factor appears to be significantly lower in IVIS+ADAS condition than the other two conditions, meaning that in IVIS+ADAS condition participants felt to be more in control than in the other two conditions; in Objective Global Variability Index, the IVIS+ADAS condition shows less variability than the No System Condition, confirming more control by drivers in the first condition.
As in DALI Stress level, BMDMW Fatigue Factor shows a significantly higher value for No System condition than ADAS condition, and Objective Global Variability Index shows significantly higher variability for No System condition than ADAS condition.

### *5.3 Discussion on CRF results*

We have decided to analyze the objective data from different levels and points of view. So, from a general point of view, computing a mean of all the track, the No System Condition, appears to be most critical, showing the highest variability.
On the other hand, analyzing the single segments, it was possible to highlight situations where the IVIS+ADAS condition appears most critical.

The conclusion is that, surely, ADAS condition presents the lowest displacement from the baseline. The No System and IVIS+ADAS conditions appear to be the most critical.

The comparison between subjective and objective data shows the following results.
DALI and BMDMW and PSA are able to show the difference among the three conditions.
In particular, it seems that DALI is more able to discriminate the single situations showing up in the single segments.
BMDMW seems to be able to take a general snapshot of the overall state of participants, as proved by the correspondence with the Objective Global variability Index.
Both DALI and BMDMW report the same trend as regards the aspects of stress, (reported in DALI as the Stress Level and in BMDMW as the Fatigue Factor): No System condition is felt as the most stressful condition.
Concerning PSA, the general effort data highlight that No systems condition and IVIS and ADAS condition seems to appear more demanding than ADAS condition.
Single factors seems to be less sensitive.
On the basis of what stated above, CRF advises to use DALI, as it is a tool able to describe the workload state of subjects; it seems sensitive to variations in workload state. It is easy to fill in and it seems more robust than BMDMW. Concerning BMDMW, the present research was the very first to use it. The preliminary results obtained are quite encouraging, but further study is needed before using it, in order to refine single factors. Regarding PSA the main problem emerged is the length to fill it in, besides the lower sensitiveness.

# 6 Experiment 2 (INRETS)

The objective of the task 226 in the AIDE project is to use various tools available for the subjective evaluation of the driver's mental workload, in order to identify the advantages and the limits of each of them according to the context and the type of system such as IVIS or ADAS.

For the experiment conducted by INRETS, 3 available tools were used in a realistic driving context, in order to define the advantages and the limits of each of them:

- the **BMDMW:** subjective evaluation of items from the driving task (see section 2.4.1)
- the **DALI** : subjective evaluation of the driver's mental workload due to the driving task (see section 2.4.2)
- the **PSA-TLX:** subjective evaluation of the performance of the task and of the effort (see section 2.4.3).

As the assessment of workload is coupled with the task difficulty (Gopher & Donchin, 1986), the general principle of the INRETS experiment was to set up experimental sessions that are varying objectively in terms of requirements for the driver, inducing then various levels of mental workload to deal with these contexts. At the end of each of these sessions, the 3 subjective evaluation tools have been applied. To summarise, the process for the experimental procedure was the following:

- To set up diversified situations varying on purpose by their level of demand: cognitive process (e.g.: to memorise the route) and perceptivo-motor process (e.g.: to run manual action following auditory, visual or tactile stimulations)
- To apply the 3 tools for each of these sessions in order to gather subjective data
- To check that the highly demand session corresponds to the highly values for the tools and to identify in which way
- To record objectives variables (driver's visual strategies, driving performance and errors, speed of the vehicle) to complete the evaluation of workload (subjective + objective)

The analysis has been conducted in order to investigate if the tools are able to reveal the differences between sessions in terms of drivers' workload, and in which way the information each of the tool brought is corresponding to the specificity of the context. In order to fulfil this general objective, the analysis has been focusing on some specific aspects linked with the tool:

- DALI
  - Investigate about the sensitivity of the various factors (visual, auditory, attention, stress…) according to the conditions of the session
  - Determine if the step of pair-wise comparison is necessary for this tool (what are the consequences of the weighing of the factors)
- PSA-TLX
  - Investigate about the values of effort and performance according to the conditions of the session (global and by driving sub tasks)
- BMDMW :
  - Investigate about the values of each item according to the conditions of the session

## *6.1 Method*

### 6.1.1 Participants

20 drivers (10 male and female) between 23 and 60 years old have participated to the study. They have a driving experience of 10 000 Km per year which would be the average mileage for a regular driver in France, and driving licence of 5 years at least. They never use a guidance system

### 6.1.2 Apparatus

The test vehicle is a scenic named MARGO. Various apparatus are located on the vehicle: miniature cameras, digital video recorder, sensors to study driver's actions, navigation computer, acquisition computer, laser scanner (see Fig. 6-1).

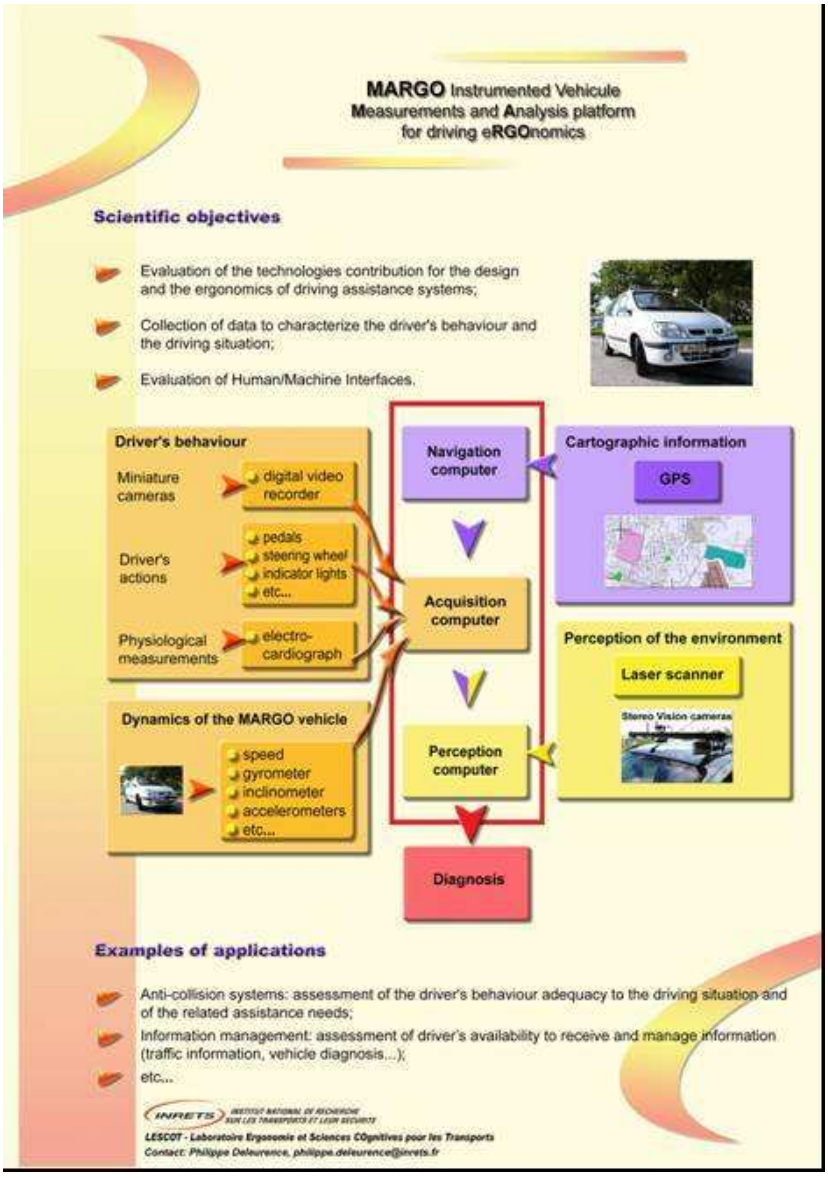



**Fig. 6-1: The in-vehicle equipments of MARGO**

The test vehicle is equipped by a guidance system. This guidance system displays visual (see Fig. 6-2) and auditory informations "Préparez-vous à prendre la 3eme sortie à gauche » and then « Prenez la 3eme sortie à gauche ».

**Fig. 6-2 : Guidance system used for INRETS experiment**

For the experiment, another on-board system has been added (see Fig. 6-3); it can display visual information (pictograms or texts) and auditory information (bips or auditory messages).

**Fig. 6-3: Screen used for INRETS experiment**

These two systems are located on the vehicle near the driver (see Fig. 6-4 )

**Fig. 6-4: Guidance system and screen integrated in the test vehicle**

The experimenter is behind the driver; he has a computer which is sending information to the on-board system (see Fig. 6-5)

**Fig. 6-5: Computer used to send information to the front screen**

With the right-hand side of the experimenter, there is a computer which records parameters of the vehicle (code of the messages sent to the driver, video time code, speed, wheel, footbrake pedal…) (see Fig. 6-6).

**Fig. 6-6: Computer recording parameters of the vehicle**

During the entire experiment, drivers were filmed; 4 cameras were used: 1 for the driver' eyes, 1 for a general view of the driver, 1 for the scene ahead and one for the HMI of the computer which is recording parameters of the vehicle. The drivers were recorded too.

### 6.1.3 Data analysis

- DALI measures

See section 2.4.2

- PSA-TLX measures

See section 2.4.3

- BMDMW

See section 2.4.1

- Workload questionnaires' evaluation: final questionnaire

See section 5.1.4

### 6.1.4 Comparison of conditions

If the objective of the experiment is to test tools and methods, then a knowledge a priori of the level of workload induced by the situation is crucial. Indeed, definition of the context will allow evaluating if the tools reflect correctly what is expected in terms of conditions and in which way the results from each subjective evaluation tool correspond to the workload deliberately induced on the driver.

Sensitivity of the DALI, BMDMW and PSA-TLX tools are evaluated through the comparison of 4 different experimental sessions varying according to the level of workload induced on the driver and as realistic as possible in a context of driving task have been tested on real road:

- **2 situations with a high task demand**:
  - **High (Context + System) HCS**: For this session, an experimental mock up which is composed by a home made system displaying several complex messages and by a navigation system had been introduced in the vehicle. In fact, while driving, the driver had to follow the route given by the guidance system and simultaneously had to run a task according to stimulations emitted by the on-board system. When considering this task, the informations to deal with are not related to the driving task and induced a manual action and a verbal answer. This session intends to

produce on purpose a high mental workload linked to ***perceptual processes***, ***decision making*** and ***motor and/or verbal output*** (detailed description in independent appendix II.1).

— **High (Context) HC**: Before the experimentation started, the driver had to consult a paper map to know the route to follow. In this case, there were no specific stimulus during the session. Then, he can stop anytime to check again the directions. The workload was linked to the ***mental representation*** of the route and to ***memorise*** it (see independent appendix II.1 for more details).

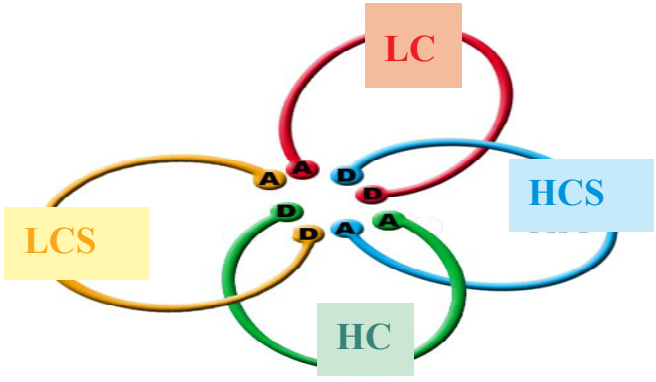
- **2 situations with a low task demand**

— **Low (Context + System) LCS:** The driver had to follow the route according to visual and auditory information given by the same guidance system as used for the HCS condition (see independent appendix II.1 for more details). The workload was linked to ***perceptual processes*** but the decision making and the mental representation/memorisation were lighter than in the previous sessions.

— **Low (Context) LC**: During the route, the experimenter gave clear and on time directions to follow. The workload was linked only with the management of the driving task, without any added activity (see independent appendix II.1 for more details).

### 6.1.5 Tasks/Systems (e.g. IVIS tasks or ADAS systems, both…)

Each tested session lasted for about 15 minutes; 10 drivers have crossed the experiment with the DALI and PSA-TLX questionnaires and 10 other drivers have used the DALI and BMDMW questionnaires
There is a turn over between the 4 sessions according to the driver. To do so, the departure point (D) and the arrival point (A) of each situation were located in the same area (see Fig. 6-7).




Routes chosen for the different experimental tested sessions:

HCS (High Context With System): urban area

HC (High Context Without System): urban area

LCS (Low Context With System): motorway area

LC (Low Context Without System): motorway area

**Fig. 6-7: Turn over between the sessions**

For example, one driver can run through HCS and go on with LC then LCS then HC and the next one run through LC, HCS, HC and LCS…

### 6.1.6 Design and Procedure

The various stages throughout the experiment can be summarised:

- Before leaving the INRETS place, the experimenter had to globally describe what the driver would have to do: to drive a car by making 4 different routes of about 15 minutes for each one from the same departure point, to complete 2 various questionnaires after each session (these questionnaires will be presented after the first session). Instructions given to drivers were to respect the Highway Code Book in agreement with the traffic rules.
  Before going to the departure point to start the experiment, the driver had to adjust the rear-mirror view and the seat. Then, he had to drive for about 10 minutes to get to better know the experimental car.

- After this familiarization run, the experimenter had to guide the driver to the departure point. Then he had to explain to the driver what he would have to do during the first session: LC, LCS, HC or HCS session because of the turn over between the sessions (see the description of each session below in the independent appendix II.1).

- After the first experimental session, the car was stopped at the same departure point and the experimenter had to present the DALI questionnaire:
  - to describe the first stage of the DALI, to explain the meaning of each factor to evaluate, and after each factor's description, let the driver complete it
  - to describe the second stage and then let the driver full it

  Then, the experimenter had to present the PSA-TLX or the BMDMW in the same way than the DALI (see independent appendix I, II2, III.2).

- When the driver had finished completing the 2 questionnaires, the experimenter had to present the second session. After this experimental tour, the driver had to fill in the 2 same questionnaires (DALI and PSA-TLX or DALI and BMDMW depending on the subject). The experiment goes on again and again until the filling of the 2 questionnaires for the fourth session.

- Then, the driver had to complete a final questionnaire to get his general feelings about the various questionnaires (see independent appendix IV.2).

### 6.1.7 Inferential statistical measures

Because of the number of tested drivers by the 3 partners INRETS, PSA and CRF, it has been decided with a common agreement to use non parametric tests for the statistical analysis; two statistical measures (common to all methods and all partners) are used:

- Wilcoxon test to compare the various sessions
- Rank correlation of Spearman to establish the correlation between the raw scores and the weighted scores for the weighted scores for the DALI

## *6.2 Results*

### 6.2.1 DALI

- DALI factors

Under this label are displayed the raw data obtained for each factor: attention, visual, auditory, tactile, temporal, interference and stress, in addition to the global score which is the computation of the average of the seven factors.
In the following graph, the values for each factor and for the global score are displayed for the 4 experimental sessions varying by their level of complexity and induce workload on the driver (see Fig. 6-8).

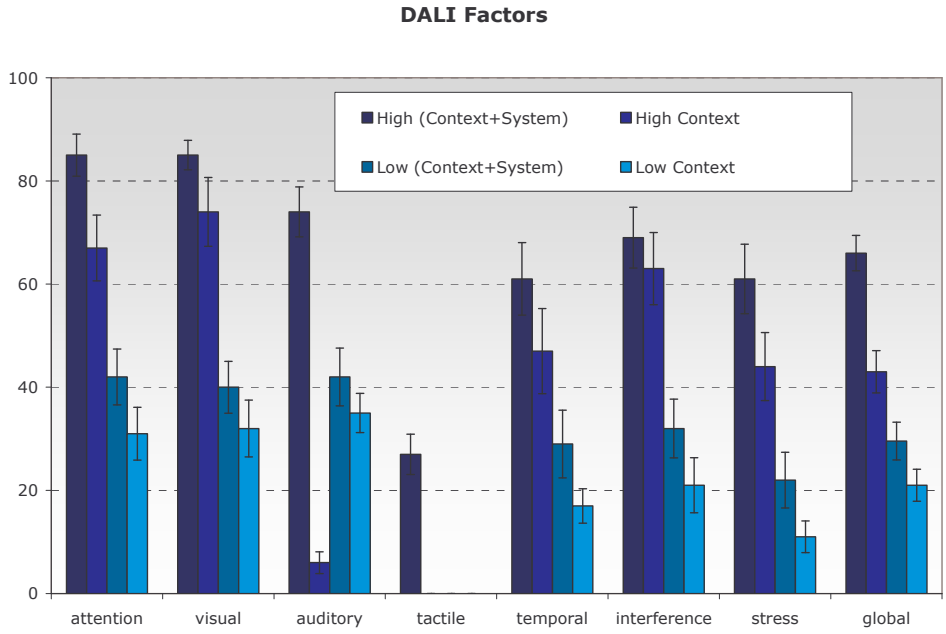



**Fig. 6-8: DALI -Values for each factor and for global score for the 4 experimental sessions**

The non parametric Wilcoxon test has been conducted in order to analyse the significance of the difference between experimental sessions (see Table 6-1).

| *Wilcoxon* | Effort of attention | Visual demand | Auditory demand | Temporal demand | Interference | Stress | Global score |
|---|---|---|---|---|---|---|---|
| LC - LCS | NS | NS | NS | NS | NS | 0.05 | 0.01 |
| LCS - HC | 0.005 | 0.001 | 0.000 | 0.05 | 0.002 | 0.025 | 0.025 |
| HCS – HC | 0.05 | NS | 0.000 | NS | NS | NS | 0.003 |
| LCS - HCS | 0.000 | 0.000 | 0.000 | 0.005 | 0.000 | 0.000 | 0.000 |
| LC - HC | 0.005 | 0.001 | 0.000 | 0.05 | 0.000 | 0.025 | 0.025 |
| LC - HCS | 0.000 | 0.000 | 0.000 | 0.005 | 0.000 | 0.000 | 0.000 |

**Table 6-1 - DALI -Non parametric Wilcoxon test used to analyse**

Global workload

There is a significant difference between the 4 experimental sessions in terms of subjective assessment of workload by the driver when looking at the DALI results (Wilcoxon, Z= 3,007, p=0,003; Z= 2,224, p=0,026, Z= 2,539, p=0,011; Z= 3,923, p<0,001).

These sessions were defined with this goal, so this result is very positive while checking the validity and the sensitivity of this tool.

The chosen sessions were varying according to various characteristics that can participate to this global workload: an analysis of the detail of the results for each factor allows to better identify and understand what are the components of this global score.

Workload linked to cognitive components

- **Attention**

There is a significant difference between the High and the Low workload sessions in terms of attentional requirements (Wilcoxon, Z= 2,840, p=0,005; Z= 3,869, p<0,001). In the High contexts, the attention required to interact with the complex on-board system is higher than the one to find his route according to the memorised information, but the difference is not that significant (Z= 1,991, p=0,047). In the Low context, there is no significant difference in terms of attention between using a guidance system and following the instructions of a co-pilot.

- **Interference**

In terms of interference, there is no significant differences between the High Context With or Without System (between HCS & HC: Wilcoxon, Z= 0,471, p=0,638). This result indicates that navigating with a paper map would be rated as interfering with the driving task as using a very complex in-vehicle system or “ergonomic mock-up” displaying several messages.

There is no significant difference between the Low Context With or Without System (between LCS & LC: Z=1,896, p=0,058). This result indicates that using a well designed in-vehicle guidance system is equivalent in terms of interference with the driving task to be guided by a human co-pilot.

Nevertheless, there is a significant difference when comparing High Context and Low Context. This difference can be explained by the type of driving area chosen for each context and the difficulties the driver encountered (driving urban area for the High Context session and motorway for the Low Context session) but it also shows that navigating with a paper map is more interfering for the driving task than using an ergonomic guidance system which is supposed to help the drivers (between HC & LCS:Z= 3,037, p=0,002, between HCS & LCS: Z= 3,662, p<0,001).

- **Stress**

There is a significant difference between most of the different types of driving contexts in terms of stress (Wilcoxon, Z= 2,382, p=0,017; Z= 2,041, p=0,041, Z= 3,880, p<0,001), with a lesser value between the High Context + System and the High Context (Wilcoxon, Z= 1,729, p=0,084). The factor stress is reflecting a global evaluation of the task constraint for the driver, and, in a coherent manner, is very low in the situation where the co-pilot is supporting the driver, a bit higher when a guidance system is fulfilling this part, much higher when the driver has to memorise his route and very high when the driver has to manage a secondary task in addition to the driving task.

Workload linked to perceptive components

- **Visual Factor**

Considering the visual demand of each session, there is a significant difference between the session with high workload High (Context + System) & High (Context) and the one with low workload Low (Context + System) & Low (Context) (Wilcoxon, Z= 3,218, p=0,001; Z= 3,95, p<0,001).

The DALI allows to show that there is no significant differences between the 2 sessions "using an on-board system displaying complex stimulations" and "using a paper map to find the route" (Wilcoxon, Z= 1,312, p=0,190; Z= 1,231, p= 0,218). There are also no significant differences between the session "to be guided by a guidance system" and "to be guided by another person". Taking into account the fact that in both situations, the driver relied on the auditory information coming from the system or from the co-pilot, it is relevant to find no significant visual workload in these two contexts.

- **Auditory Factor**

Considering the auditory demand of each session, a very low value of workload is displayed in the situation where the driver has to memorise his route with a paper map and to find his way based upon the road directions in comparison with the 3 other situations (significant difference (Wilcoxon, Z= 3,954, p<0,001; Z= 3,771, p<0,001; Z= 3,804, p<0,001). Indeed, in this case, even if the general workload of the situation appeared to be high, the DALI results show that the auditory demand is not involved in this workload.

Furthermore, there is no significant difference between the situation "using a guidance system" and following instructions from a co-pilot, showing that the auditory messages coming from the on-board system did not induce a noticeable workload by the driver (Wilcoxon, Z= 1,144, p=0,253).

- **Tactile Factor**

Implementation of vibrations in the seat of the vehicle was a first approach to define if the driver was able to detect this kind of "unusual" stimulus with accuracy, and if this stimulus was inducing workload. The tactile stimulations were quite well detected and induced a light workload in comparison with situations where this stimulation was non-existence (Wilcoxon, Z= 3,703, p<0,001). Nevertheless, this workload is far less important than the one induced by auditory and by visual stimulations for the same session.

Workload linked to temporal components

Like for the **global score**, for the **stress** and for the **attention**, the **temporal** demand is highly different in relation to the type of session (Wilcoxon, Z= 1,118, p=0,264; Z= 1,556, p=0,120, (Wilcoxon, Z= 2,116, p<0,034; Z= 2,843, p=0,004). Indeed, like the other 3 factors, this factor is revealing a global estimation of the cost of the task. As driving task is under time constraint, it is then not surprising to have a workload value in terms of timing closely linked to the level of the task complexity.

## Summary of main results from the DALI factors

The values of the DALI factors showed the significant difference between the 4 experimental sessions, defined a priori on purpose with an increased level of workload for the driver: this tool allowed in a quick and reliable way to identify the global workload of a given context, and to bring additional precision about the level of load for the vision, the audition, the stress, the attention components for each of these driving contexts.

The values of driver's load (visual, auditory and attentional demands) are not significantly different in the context « using a regular guidance system implemented in the vehicle" and the context of a "co-pilot giving verbal guidance instructions to the driver". These results showed that the implemented system in this case was correctly designed in terms of visual and auditory messages (timing, loudness, content) and is not inducing noticeable attentional requirement in terms of management of a secondary task. Nevertheless, the DALI results showed that there is a slightly higher level of stress while using the system in comparison with relying on the human co-pilot. These results showed that this tool is sensitive to various aspects of the driving task, and can then support the design process by identifying which part of the task was heavier for the driver. In this specific case, the conclusion would be that the guidance system is correctly designed, but that its use requires a phase of familiarisation for the driver to be fully comfortable with it.

The values of driver's load in terms of interference are no significantly different between the High Context With or Without System, indicating that "navigating with a paper map" would be rated by the drivers as interfering with the driving task in the same way as "using simultaneously a complex ergonomic mock-up" (this experimental ergonomic mock up is composed by a home made system displaying several complex messages and by a navigation system, and intends to produce on purpose a high mental workload for the user).

The values of driver's load in terms of interference are no significantly different between the Low Context With or Without System, indicating that using a well designed "in-vehicle guidance system" is equivalent in terms of interference with the driving task to be guided by a "human co-pilot".

Nevertheless, there is a significant difference when comparing High Context and Low Context, indicating, among other things, that "navigating with a paper map" is more interfering for the driving task than "using a guidance system".

- DALI Weighted

Under this label are displayed the data for each factor (attention**, visual, auditory, tactile, temporal, interference and stress** & **global score**) weighted according to the results of the pair wise comparison between factors (see Independent appendix II.2 for the list of pair wise possibilities). This computation is also conducted in the NASA-TLX. This process conducted to two sets of data: values of factors (paragraph above) and values of weighted factors (see Fig. 6-9).

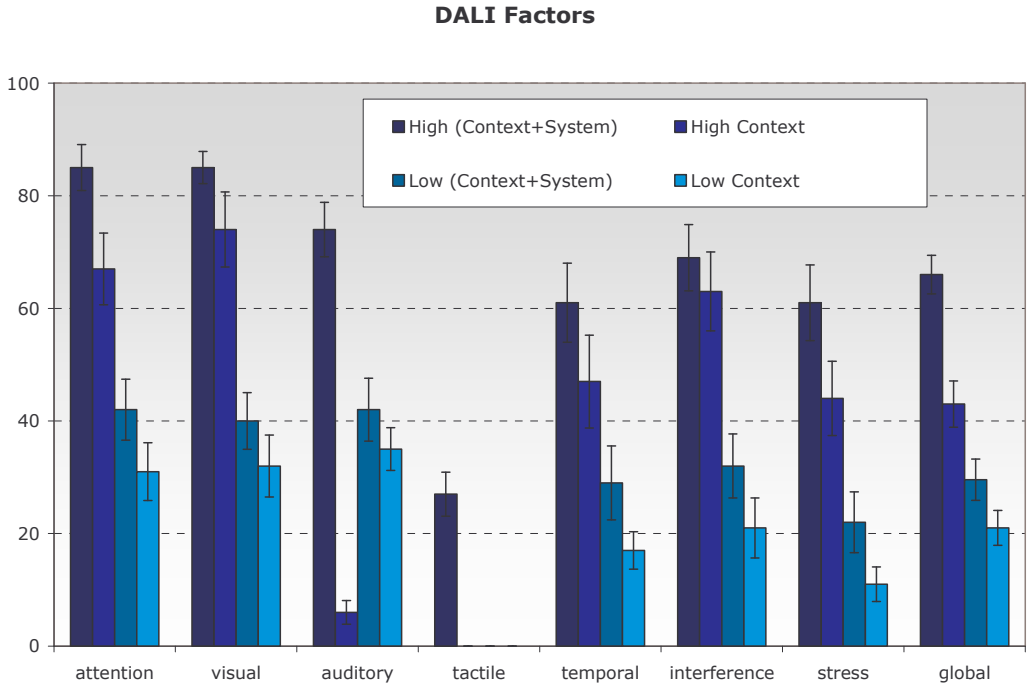



**Fig. 6-9: DALI -Values for each weighted factor and for global score for the 4 experimental sessions**

The non parametric Wilcoxon test has been conducted in order to analyse the significance of the difference between experimental sessions (see Table 6-2).

| *Wilcoxon* | Effort of attention | Visual demand | Auditory demand | Tactil demand | Temporal demand | Interference | Stress | Global score |
|---|---|---|---|---|---|---|---|---|
| LC - LCS | NS | NS | NS | 0.01 | NS | NS | NS | NS |
| LCS - HC | 0.01 | 0.025 | 0.000 | 0.01 | NS | 0.000 | 0.01 | 0.01 |
| HCS – HC | NS | NS | 0.000 | 0.01 | NS | NS | NS | NS |
| LCS - HCS | 0.001 | 0.001 | 0.000 | 0.01 | 0.025 | 0.000 | 0.01 | 0.01 |
| LC - HC | 0.01 | 0.025 | 0.000 | 0.01 | 0.01 | 0.000 | 0.01 | 0.01 |
| LC - HCS | 0.001 | 0.001 | 0.000 | 0.01 | 0.001 | 0.000 | 0.01 | 0.01 |

**Table 6-2 - DALI -Non parametric Wilcoxon test used to analyse weighted factors**

Globally, the main tendencies obtained when considering DALI Factors and DALI Weighted Factors results are close, and follow the same logic in terms of rating of the various contexts.

One of the differences is that the weighting of the values led to less significant differences after computation of the Wilcoxon test, in comparison with the Factors values. There are 4 cases, circles in red on the table above, that are not displaying significant difference with Weighted Factors values while the differences were significant with Factors values. In fact, it seems that the fact to ask drivers to compare two by two the factors induced, in one way or another, more spreading out of the scores.

Subjective evaluation after running an experiment is a very subtle process, as it requires for the driver to rely on his memory to judge what was the more costly for him. It is why it is so important to have light tool, quickly process, so the memory is still fresh at the end of the questionnaire.

One hypothesis is that asking drivers to compare factors after rating factors on the scale, could induced an additional task leading for some of them to confusion rather than making the rating more precise.

- Comparison between the Factors Values and the Weighted Values

In order to check the correlation between these two set of values, a test of Spearman has been conducted (see Table 6-3).

| Spearman (correlation of weighted and non-weighted factors) | Effort of attention | Visual demand | Auditory demand | Tactil demand | Temporal demand | Interference | Stress | Global score |
|---|---|---|---|---|---|---|---|---|
| HCS | 0.01 | 0.01 | 0.01 | No tactile | 0.01 | 0.01 | 0.01 | 0.05 |
| HC | 0.01 | NS | 0.01 | No tactile | 0.01 | 0.01 | 0.01 | 0.05 |
| LCS | 0.01 | 0.01 | 0.01 | No tactile | 0.01 | 0.01 | 0.01 | 0.01 |
| LC | 0.01 | 0.05 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.05 |

**Table 6-3 - DALI -Spearman test to analyse the correlation between factors and weighted factors**

All the values are highly correlated, except the values for visual workload in the case of High Context without System, labelled NS on the table above. In this specific case, the situation "driving with a paper map" seemed to be rated by the driver, after pair wise comparison, in a different way than when rating directly the Factors on the first phase. According to observation of the driver's behaviour made by the experimenter during the session, there were not that many reasons about this difference. Our hypothesis is that this difference would be more an artefact.

According to these results, it appears that the computation of the pair wise comparison would not bring any added value to the Factors rating. On the contrary, the fact to suppress this stage would have the advantage to make even lighter to process (data gathering, data processing) the DALI as a tool for the workload subjective evaluation.

- Comparison between young and older drivers

According to Fig. 6-10, the data of the DALI weighted factors for the global score values have been split into two groups of drivers: the "young" one, n=13 from 21 to 35 years old, and the "old" one, n= 7, from 43 to 62 years old. The second group is not very old, but

corresponds to a population presenting evolutive presbyopia. Previous studies (Pauzié, 1992) have shown that this age group, around 50 years old, has a vision the less efficiently corrected, especially the short distance and middle distance vision. Indeed, the visual ability is quickly changing at this age so the people are not necessarily aware about these changes. People over 60 years old have in average better correction of their vision. This is important as most of the screens implemented in the vehicle are located at this type of critical distance for presbyopia persons.

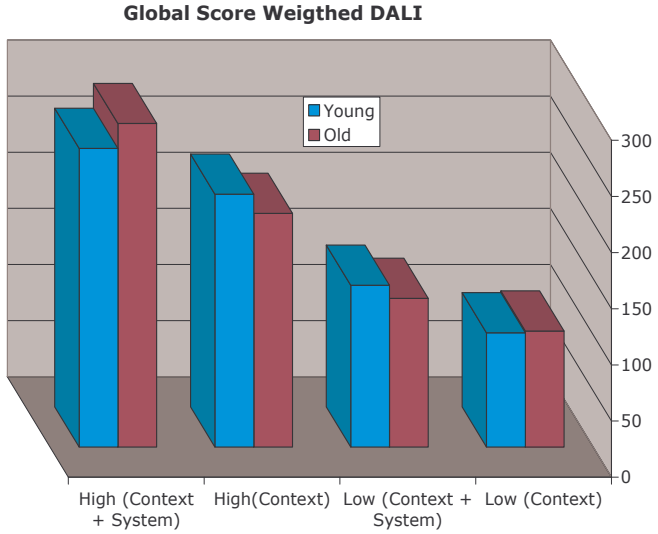



**Fig. 6-10: DALI -Comparison between young and older drivers**

Globally, there were no differences of workload in relation to the type of session whatever the age of the driver.

## Summary of main results from the DALI weighted factors

Globally, the main tendencies obtained when considering DALI Factors and DALI Weighted Factors results are close, and follow the same logic in terms of rating of the various contexts. Computation of the correlation between the Weighted Factors values and the Factors values showed that these results are strongly correlated.

According to these results, it appears that the computation of the pair wise comparison would not bring any added value to the Factors rating. On the contrary, the fact to suppress this stage would have the advantage to make even lighter to process (data gathering, data processing) the DALI as a tool for the workload subjective evaluation.

### 6.2.2 PSA-TLX results

The results of the PSA-TLX questionnaire have been computed according to the available process defined by PSA, using pre-defined excel files. The template was pre-defined to an experiment protocol with 3 sessions, so it has been modified in order to be able to be used for this experiment with 4 sessions. These excel files allowed to obtain numerous graphs which are all presented in the Independent appendix II.5 and II.6:

The results concerning Global Performance and Effort, Performance and Effort for each sub-task, Performance and Effort for each session with a distribution of scores for each subject, Global compromise Performance/Effort, Comparison of compromise Performance/Effort for each sub-task of driving are discussed in this paragraph.

- **Global Performance and Global Effort**

In the PSA-TLX tool, there is a global score for two aspects of the driving task: subjective evaluation regarding the performance of the driving, and subjective evaluation regarding the effort made during the session. These evaluations are rated in comparison with the usual normal driving as a reference situation.

The Fig. 6-11 displays the results for the Global Performance according to the 4 experimental sessions.

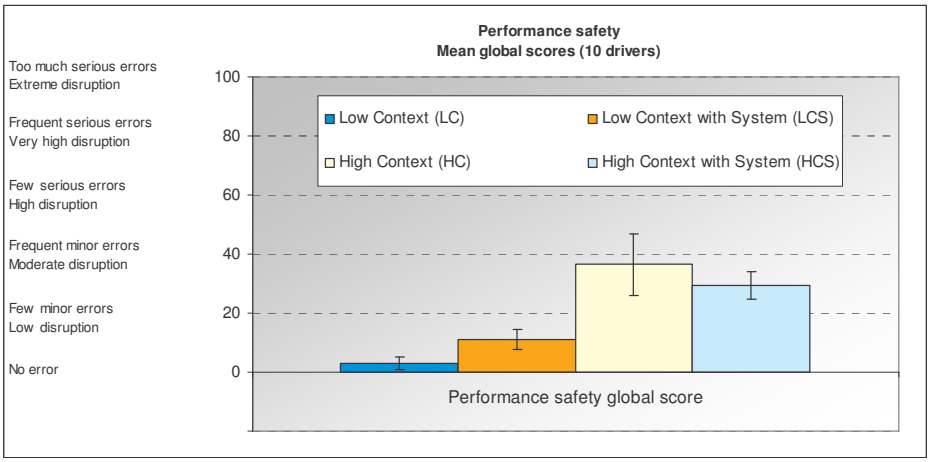



**Fig. 6-11: PSA-TLX -Mean global scores for performance safety**

| Global score Performance | Conditions LC - LCS | Conditions LCS - HC | Conditions HCS - HC | Conditions LCS - HCS | Conditions LC - HC | Conditions LC - HCS |
|---|---|---|---|---|---|---|
| Wilcoxon | **NS** | **0.025** | **NS** | **0.025** | **0.025** | **0.025** |

**Table 6-4 - PSA-TLX -Wilcoxon test used to compare global scores performance**

The difference between LC-LCS (“human co=pilot” and “guidance system”) and HCS-HC (“complex ergonomic mock-up” and “use of paper map”) did not happen to be significant according to the statistical analysis (see Table 6-4). The DALI values indicated that in the first case, there were no significant differences in terms of visual, auditory, attention, interference and temporal demand, but a difference in terms of stress, due certainly to the fact that the driver was not familiarized with the guidance system.

In the case of evaluated performance, the situation "finding the route using a paper map" – HC – was rated as the one inducing the higher workload. Indeed, wavering manoeuvres were common during this session, and the driver remembered clearly these phases when filling in the questionnaire at the end of this session. Objective analysis of the data performance from the experimental vehicle will indicate if there was an objective lowering of the performance in this situation, or if it was more a question of subjective feeling from the driver after running the session.

**Performance safety and criteria**

| | Mean Values | Values | Analysis* |
|---|---|---|---|
| **Low Context** | 3,00 | <5 | Driver's performance is optimal and there is no risk |
| **Low Context with System** | 11 | [5 - 20] | Driver's performance is acceptable and there is no notable deterioration |
| **High Context** | 36,50 | [25 - 40] | The performance is moderately damaged |
| **High Context System** | 29,50 | [25 - 40] | The performance is quietly damaged |

**Table 6-5 - PSA-TLX -Analysis of mean values for each experimental session for performance safety**

According to the report "Description of procedure to analyze the PSA-TLX, AIDE internal report, 2005," (see * Appendix III.3 for a complete description of the analysis), the safety criteria for performance is 55. The Table 6-5 shows that the values for LC, LCS, HC and HCS are lower than 55:

- In the Low Context session, the performance stays optimal (3); there is no error and no disruption. For the drivers of this experiment, there is no modification of their usual driving activity.

- In the Low Context with System session, there is no appreciable change of driver's performance; a few number of errors have been done but they are not serious at all; the level of performance stays very acceptable.

- In the High Context session where the driver has to find his route by using a paper map, frequent errors have been done but they are not serious; the performance of drivers is moderately damaged but stays acceptable.

- In the High Context System session where the driver was using a highly demanding in-vehicle ergonomic mock-up, driver's performance is quietly or moderately damaged; minor errors and minor disruption have been done. The level of performance is acceptable but not optimal.

The use of the on-board guidance system and the guidance by a co-pilot was rated as safe in terms of Performance. However, to see if using a complex in-vehicle system and remembering the route on a paper map are rated as "unsafe situations", it is important to analyse the effort felt by the drivers.

The following graph displays the results for the Global Effort according to the 4 experimental sessions.

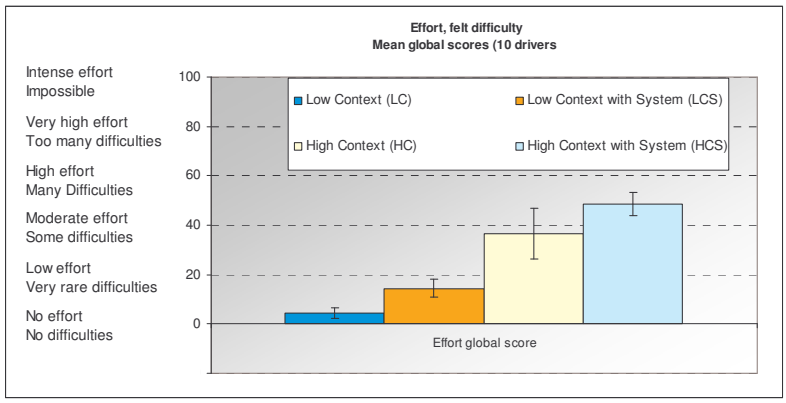



**Fig. 6-12: PSA-TLX -Mean global scores for effort felt**

In the case of effort, the PSA-TLX values are organized in a similar way as the values obtained with the global score of the DALI, that is to say an increased graduated workload from Low Context to High Context with System. It could be the effort linked to perception and/or to cognition (see Table 6-6).

| Global score Performance | Conditions LC - LCS | Conditions LCS - HC | Conditions HCS - HC | Conditions LCS - HCS | Conditions LC - HC | Conditions LC - HCS |
|---|---|---|---|---|---|---|
| Wilcoxon | **NS** | **0.025** | **NS** | **0.01** | **0.025** | **0.01** |

**Table 6-6 - PSA-TLX -Wilcoxon test used to compare global scores effort**

Nevertheless, like for the performance, the difference between LC-LCS ("human co-pilot" and "guidance system") and HCS-HC ("complex ergonomic mock-up" and "use of paper map") did not happen to be significant according to the statistical analysis. The global score of the DALI showed a significant difference between "complex ergonomic mock-up" and "use of paper map", due to attentional demand (more attention for using the complex ergonomic mock-up) and auditory demand (no auditory demand for using the paper map). The values of the DALI confirmed that these two situations were demanding in a similar way in terms of visual, temporal, stress and interference.

**Safety criteria**

| | Mean Values | Values | Analysis* |
|---|---|---|---|
| **Low Context** | 4,50 | <5 | There is no effort and no difficulties. |
| **Low Context System** | 14,50 | [5 - 20] | Driver's effort is low; the task is not demanding; difficulties are very rare. |
| **High Context** | 36,50 | [25 - 40] | The effort is moderate; few difficulties. |
| **High Context System** | 48,50 | [45 - 60] | The effort is high; many difficulties. |

**Table 6-7 : PSA-TLX -Analysis of mean values for each experimental sessions for effort felt**

According to the report "Description of procedure to analyze the PSA-TLX, AIDE internal report, 2005" (see * Appendix III.3 for a complete description of the analysis), the safety

criteria for effort is 55. The table above shows that the values for LC, LCS, HC and HCS are lower than 55:

- For the LC session (reference situation with a human co-pilot), driver's effort is non-existent or very low. According to the comments of drivers for this session, driving activity was very similar to their usual driving activity. The effort required is the same as usually and is very acceptable.

- For the LCS session, the use of the on-board guidance system don't required notable effort too (low effort and very rare difficulties). In fact, according to the comments of drivers, driving activity was not demanding and very easy with the help of the guidance system. Drivers have the feeling to drive as well as usually.

- For the HC session, the effort felt is moderate and drivers had some difficulties; However, the level of effort required for driving activity is acceptable.

- For the situation HCS, the effort felt is high and drivers had many difficulties. The driving activity is demanding but the level of effort required is not critical.

### Compromise between global Performance and global Effort

In the analysis of this tool, it is possible to compute and to represent the compromise between Effort and Performance. The objective is to put in balance the results got from the estimated Performance values and the results got from the evaluated Effort values in relation to the session and to the driver.

Indeed, there is not necessarily a high correlation between the assessed cost for a task and the observed performance while running this task. The driver can have the strategy to increase his effort in order to maintain at the same performance level in case of an increased complexity of the task. In this case, values of the performance stay stable while effort increases. This is especially valid when considering objective performance measured through recorded driving errors, observation of wavering manoeuvres.

In the case of subjective performance evaluated by the driver, the issue is a bit different as there can be a gap between the real performance and the way it is stated by the driver. At this stage of the analysis, there is an uncertainty about the correlation between the objective performance, based upon measured of driver's behaviour and vehicle parameters, and subjective performance, evaluated by the driver at the end of the session. There is usually a strong stress from the society about performance in general, and driving performance in particular. So, the hypothesis can be made that the driver will have a tendency to over-estimate his driving performance. Or, at least, not to have the possibility to clearly evaluate it (for instance, subjective evaluation of the performance for the control of lateral motion and longitudinal motion in a normal driving is not easy to do. The action to maintain the vehicle in the lane is an automatic process for experienced driver and the quality of this action is not necessarily available as a conscious process for the driver). Further analysis of the data would allow to check this hypothesis.

This question of overestimation by the driver is certainly less strong for subjective evaluation of Effort, where, in comparison with Performance, pressure from the society is less important in this case.

The analysis made available in the computation of the PSA-TLX allowed to compute the compromise Performance / Effort for each sub-task of the driving activity and for each driver. In order to have a global overview as a first step of the analysis, we compute also the compromise of the global values for Performance and Effort.

Compromise for Global Performance and Global Effort

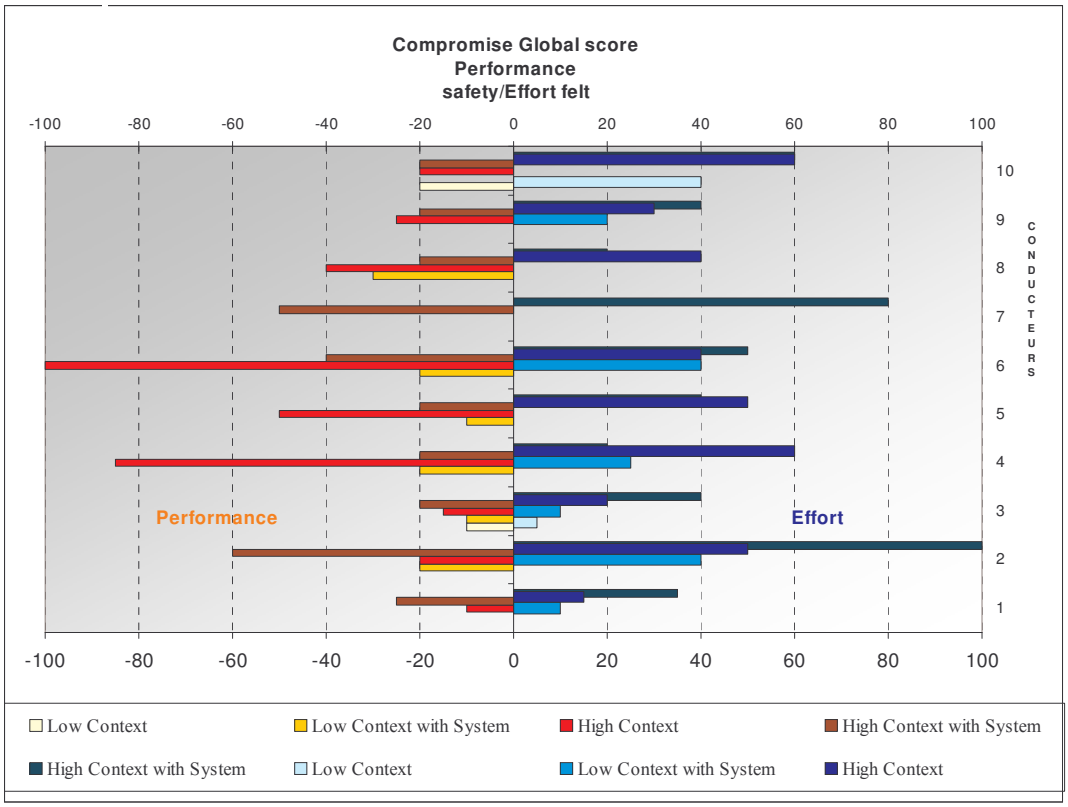



**Fig. 6-13: PSA-TLX -Compromise for global performance and global effort**

The proposed graph for compromise computation is difficult to interpret, partly because the display is made subject by subject. In order to overcome this difficulty, we propose the same graph with the computation of the average for each session (displayed horizontally, rather than vertically) (see Fig. 6-14).

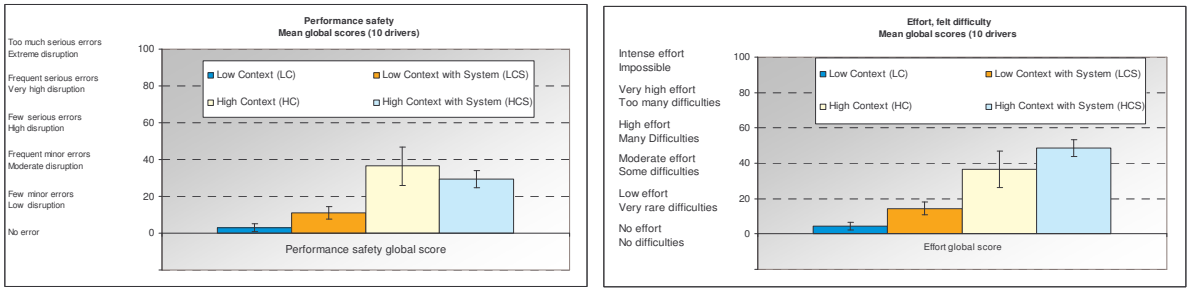



**Fig. 6-14: PSA-TLX -mean global scores for global performance and effort felt**

For the three sessions, “co-pilot”, “guidance system” and “using a paper map”, Performance and Effort seemed to be correlated with an equilibrated compromise between these two criteria.

It is not the case for the session “using a complex ergonomic mock-up” HCS where the Performance evaluated by the driver presented a level of safety “better” than with the paper map, while Effort presented a higher level. One possible interpretation can be that drivers managed to use the complex in-vehicle mock-up by increasing their effort in order not to penalize the driving task performance.

To complete this approach, it would be interesting to compute the correlation between values of the Performance and values of Effort for each session.

- Sub-tasks Performance

The global score for Performance presented above is the global estimation made by the driver of his performance (see Fig. 6-15). Then, in a second phase of the questionnaire, the driver has to rate a set of sub-tasks describing several aspects of the driving task related to performance.

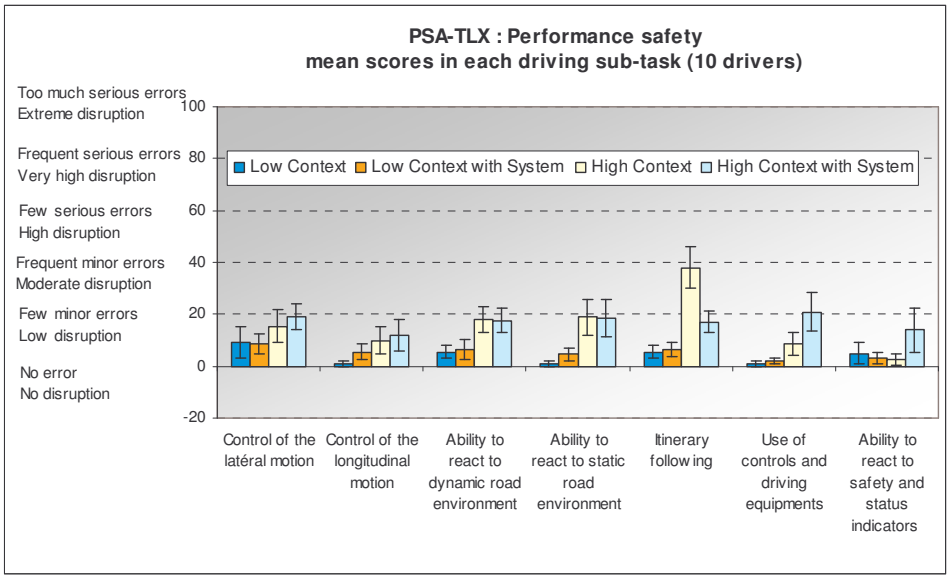



**Fig. 6-15: PSA-TLX -mean scores of performance safety for each driving sub-task**

| *Wilcoxon* | Lateral | Longitudinal | Dynamic road environment | Static road environment | Itinerary | Controls and driving equipments | Safety and status indicators |
|---|---|---|---|---|---|---|---|
| LC / LCS | NS | NS | NS | NS | NS | NS | NS |
| LCS -/HC | NS | NS | 0.05 | NS | 0.025 | NS | NS |
| HCS / HC | NS | NS | NS | NS | 0.05 | NS | NS |
| LCS / HCS | NS | NS | NS | NS | NS | 0.025 | NS |
| LC / HC | NS | NS | 0.05 | 0.05 | 0.025 | NS | NS |
| LC -/HCS | NS | NS | NS | 0.05 | 0.05 | 0.025 | NS |

**Table 6-8 - PSA-TLX -Wilcoxon test used to compare experimental sessions for performance for each sub-task**

The Table 6-8 presents a comparison of the different experimental sessions with the Wilcoxon test:

- For itinerary following, there is no significant difference between the sessions "co-pilot" and "guidance system" and between the session "guidance system" and "complex ergonomic mock-up".

- There are no significant differences for evaluation of lateral motion and longitudinal motion, and for safety and status indicators, whatever the session.

- For the ability "to react to dynamic road environment", there is a significant difference for the situations "using a paper map" and "co-pilot" on one side, and "guidance system" on the other side. "Using a paper map" conducted the drivers to evaluate their performance as not good in comparison with the two other contexts.

- For the ability "to react to static road environment", there is a significant difference for the situations "co-pilot" and "using a paper map" on one side, and "using a complex ergonomic mock-up" on the other side. To follow the instructions of a co-pilot lead the drivers to evaluate their reaction to the static road environment better in comparison with the two other contexts.

- For the "use of controls and driving equipments", there is a significant difference for the situations "using a complex ergonomic mock-up" and "using a guidance system" on one side, and "co-pilot" on the other side. Using a complex in-vehicle system leads the drivers have to evaluate their use of controls and driving equipments worst in comparison with the two other contexts.

**Safety criteria**

| mean score | Control of the lateral motion | Control of the longitudinal motion | Ability to react to dynamic road environment | Ability to react to static road environment | Itinerary following | Use of controls and driving equipments | Ability to react to safety and status indicators |
|---|---|---|---|---|---|---|---|
| Low Context | 9,00 | 1,00 | 5,50 | 1,00 | 5,50 | 1,00 | 5,00 |
| Low Context System | 8,50 | 5,50 | 6,50 | 4,50 | 6,50 | 2,00 | 3,00 |
| High Context | 15,50 | 10,00 | 18,00 | 19,00 | 38,00 | 8,50 | 2,50 |
| High Context System | 19,00 | 12,00 | 17,50 | 18,50 | 17,00 | 21,00 | 14,00 |

**Table 6-9 - PSA-TLX -Analysis of mean values for each experimental session for performance got for each sub-task**

According to the report "Description of procedure to analyze the PSA-TLX, AIDE internal report, 2005" (see * Appendix III.3 for a complete description of the analysis), when considering the safety threshold for performance stated in the PSA-TLX, the level of performance is acceptable for most of driving sub-task whatever the experimental sessions: only few minor errors have been done and the disruption is low.

However, in the situation where the driver is "using a paper map" – High Context – for the sub-task "performance in itinerary following", the mean value is 38. In this case, frequent errors have been done but they are not serious; driver's performance is moderately damaged; the level of performance is not optimal but stays acceptable.

- Sub-tasks Effort

The global score for Effort presented above is the global estimation made by the driver of his effort when answering to a first question. Then, in a second phase, the driver has to rate a set of sub-tasks describing several aspects of the driving task related to effort.

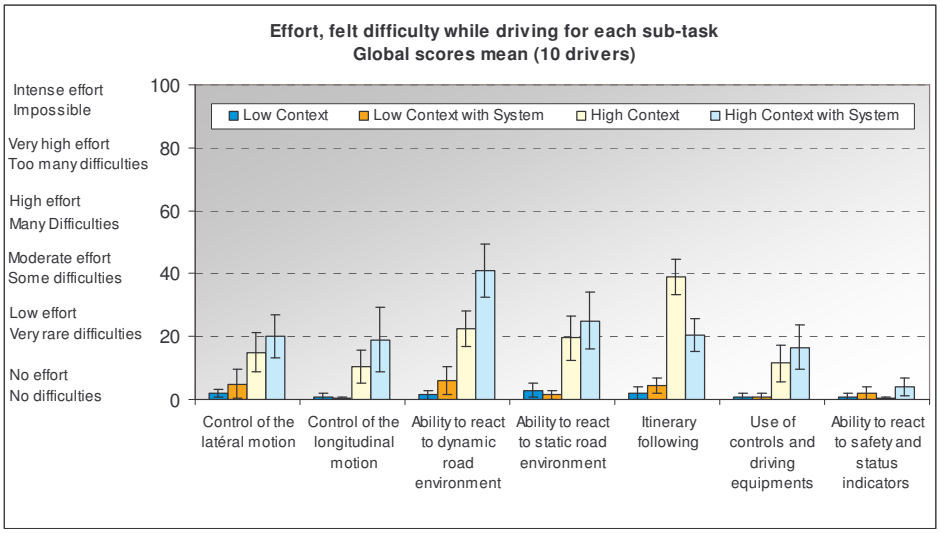



**Fig. 6-16: PSA-TLX -Mean scores of effort felt for each driving sub-task**

| *Wilcoxon* | Lateral | Longitudinal | Dynamic road environment | Static road environment | Itinerary | Controls and driving equipments | Safety and status indicators |
|---|---|---|---|---|---|---|---|
| LC/LCS | NS | NS | NS | NS | NS | NS | NS |
| LCS /HC | NS | NS | 0.05 | 0.025 | 0.01 | NS | NS |
| HCS / HC | NS | NS | 0.05 | NS | NS | NS | NS |
| LCS / HCS | NS | NS | 0.025 | 0.05 | 0.05 | 0.05 | NS |
| LC / HC | NS | NS | 0.025 | NS | 0.01 | NS | NS |
| LC / HCS | 0.025 | NS | 0.025 | 0.05 | 0.025 | 0.05 | NS |

**Table 6-10 - PSA-TLX -Wilcoxon test used to compare experimental sessions for effort for each sub-task**

**Safety criteria**

| mean score | Control of the latéral motion | Control of the longitudinal | Ability to react to dynamic road environment | Ability to react to static road environment | Itinerary followin g | Use of controls and driving | Ability to react to safety and status |
|---|---|---|---|---|---|---|---|
| Low Context | 2,00 | 1,00 | 1,50 | 3,00 | 2,00 | 1,00 | 1,00 |
| Low Context | 5,00 | 0,50 | 6,00 | 1,50 | 4,50 | 1,00 | 2,00 |
| High Context | 15,00 | 10,50 | 22,50 | 19,50 | 39,00 | 11,50 | 0,50 |
| High Context | 20,00 | 19,00 | 41,00 | 25,00 | 20,50 | 16,50 | 4,00 |

**Table 6-11 - PSA-TLX -Analysis of mean values for each experimental session for effort for each sub-task**

When considering the report of PSA described in Appendix III.3, the value of the safety threshold for effort stated is 55. The values of the Table 6-11 showed that the level of effort is acceptable for most of driving sub-tasks, whatever the experimental sessions: the effort required to drive is low and the difficulties are rare. According to the comments of drivers,

driving in these different contexts was very similar to their usual driving activity. The effort required is the same as usually and is acceptable.

Considering the HC session and the driving sub-task "Itinerary following", the effort required is moderate and drivers have some difficulties. The level of effort is acceptable considering the safety threshold for effort (55).

For the HCS session and the driving sub-task"the ability to react to dynamic road environment", the mean values (41) show that the effort required is moderate and that the drivers have some difficulties to drive. However, the level of effort doesn't exceed the safety threshold for effort and is acceptable.

- Management of performance

It is proposed to conduct the analysis by computing the relative ratio of each sub-task component. The following graph displays the values for each experimental session for Performance and for Effort.

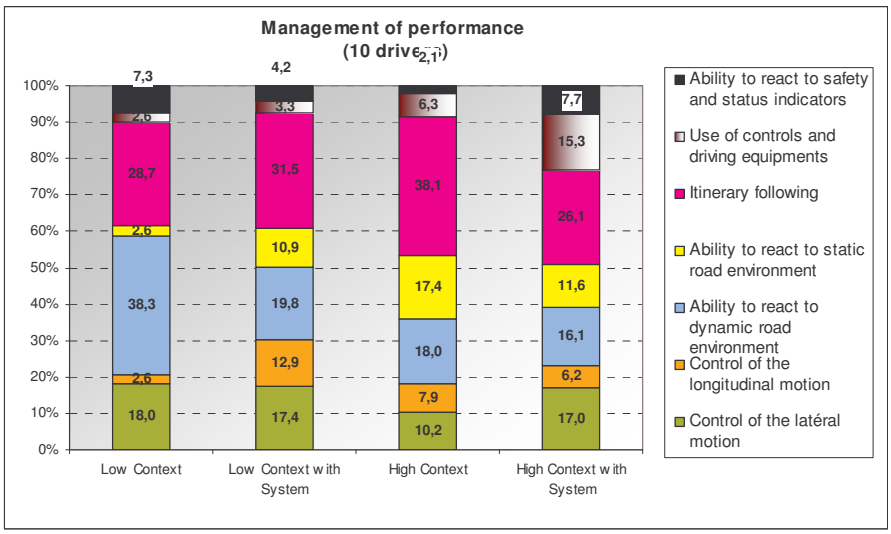



**Fig. 6-17: PSA-TLX -Management of performance**

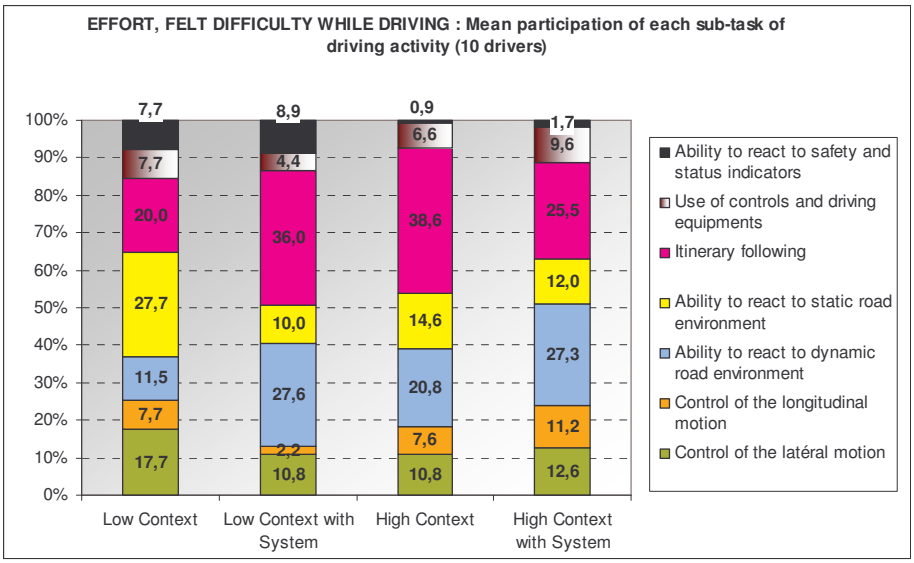



**Fig. 6-18: PSA-TLX -Managment of effort**

Relative comparison of each sub-task for a given situation

This approach allows evaluating the part taken by each sub-task in the overall subjective Performance and Effort for a given experimental session. For example, when considering the session "guidance with paper map", session HC, the item "itinerary following" is the one that participates to the highest value of evaluated Performance and for Effort in comparison with the other sub-tasks.

Comparison of situations

This approach is not valid to compare experimental sessions one with the other, as the global value of subjective Performance and Effort is different in each case. It is an artificial process to bring them to the same scale of 100%.
For example, considering the comparison between the LCS situation and the HCS situation for "ability to react to dynamic road environment": when computing relative ratio of each component, this item participated for 27,6% in LCS and for 27,3% in HCS. Nevertheless, it is not possible to consider that the workload induced by this activity is equivalent in the two situations. The graph displaying the average values indicates that this item was rated as much higher in the HCS than in the LCS experimental session.

So, this type of graph can be useful to identify the relative participation of each driving sub-task for one session on a scale of 100%. However, when considering a specific driving sub-task, it is tricky to compare performance and effort between two sessions for this sub task as they do not have the same initial value and they have been both brought back to 100%.

- Compromise Performance / Effort in driving Sub-tasks

Compromise Performance / Effort for the Lateral control

The report for the analysis of Compromise Performance and Effort is recommended to compute all the sets of compromise for all the 7 sub-tasks, driver by driver (see Fig. 6-19).

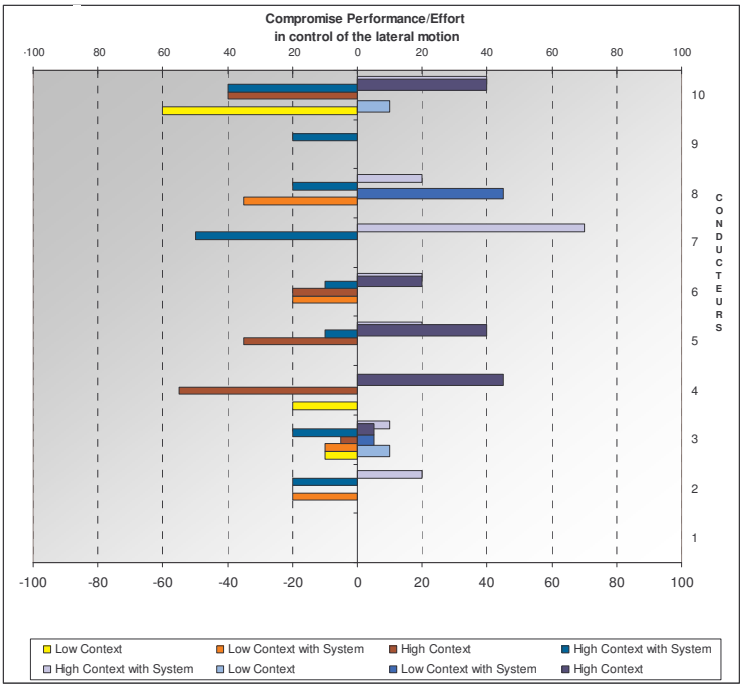



**Fig. 6-19: PSA-TLX -Compromise between performance and effort in control of the lateral motion**

The example above presents the results of the Lateral Motion, with Performance on the left and Effort on the right for each of the 4 experimental sessions. The list of results of compromise computations for each sub-tasks are displayed in the independent appendix II.5.

The objective is to gather the overall available information, but it is difficult to interpret the data for the four experimental sessions displayed on the same graph.
The proposition was to compute the Compromise and to display only two experimental sessions on the same graph, rather than the four. The problem is then the number of possible combinations while using the 7 sub-tasks (42 graphs and tables to analyse), and the resulting effort and time consuming of the analysis.

In order to simplify, we propose to conduct the process for the global value of Performance and Effort allowing having a global overview of each driving condition.
The following graph is displaying the values of the compromise for Performance and Effort by comparing two sessions: LCS (“use of guidance system”) and HC (“use or paper map”). The other cases are gathered in the Independent appendix II. 5.

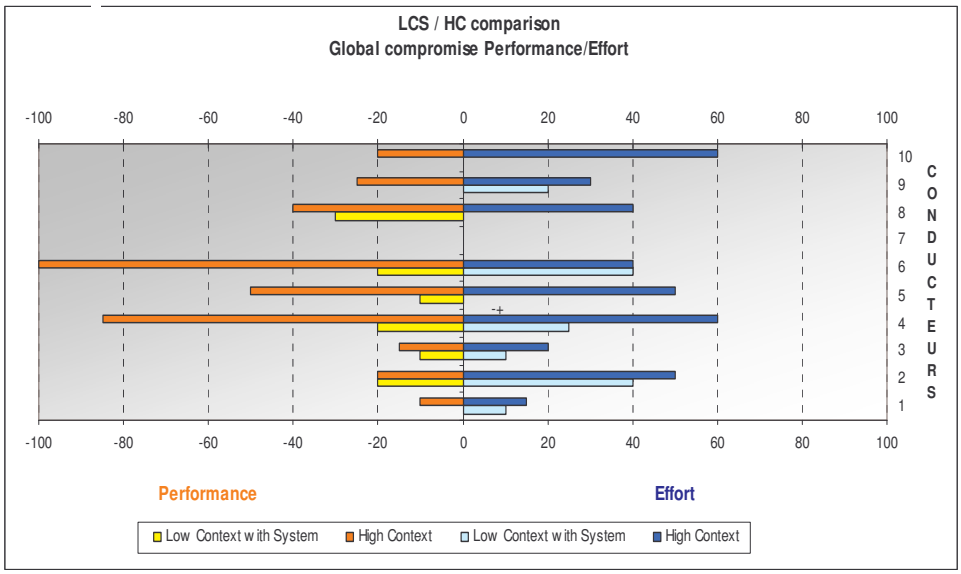



**Fig. 6-20: PSA-TLX -Compromise between global performance and global effort for LCS and HC conditions**

In order to analyse the data, the proposed process is to check for each driver the positive or negative evolution of the Effort and Performance value, which allowed to give a note to the compromise from "Optimal" to "very unfavourable".

| Effort felt | | Performance safety | | | |
|---|---|---|---|---|---|
| Evolution | Level | Evolution | Level | Drivers | Compromise Performance/Effort |
| = | 0 | = | 0 | 7 | Optimal |
| ↗ | 0<E<20 | ↘ | 0<P<20 | 1 | Favourable |
| ↗ | ≥40 | = | 20 | 2 | Unfavourable |
| ↗ | 20 | ↘ | 0<P<20 | 3 | Favourable |
| ↗ | 60 | ↘ | ≥80 | 4 | Very unfavourable |
| ↗ | 40≤E≤60 | ↘ | 40≤P≤60 | 5 | Very unfavourable |
| = | 40 | ↘ | 100 | 6 | Very unfavourable |
| ↗ | 40 | ↘ | 40 | 8 | Very unfavourable |
| ↗ | 20≤E≤40 | ↘ | 20≤P≤40 | 9 | Unfavourable |
| ↗ | 60 | ↘ | 20 | 10 | Very unfavourable |

**Table 6-12 - PSA-TLX -Analyse of global compromise Performance/Effort for LCS and HC conditions**

Summary of the global compromise Effort/Performance in the sample

| Efficiency of the compromise | Number of drivers |
|---|---|
| Optimal | 1 |
| Very favourable | 0 |
| favourable | 2 |
| Unfavourable | 2 |
| Very Unfavourable | 5 |
| Total | 10 |

**Table 6-13 - PSA-TLX -Summary of global compromise Performance/Effort for LCS and HC conditions**

For 70% of drivers (7/10), the compromise between Performance and Effort is unfavourable and even very unfavourable when drivers are driving and following the route with the help of a map. These results display the benefit to use this navigation system.
For 30% of drivers (3/10), the compromise is favourable and even optimal. The use of a navigation system is not as efficient as a map because of the routes that the drivers know.

## Summary of main results from PSA-TLX Effort and Performance

The differences between most of the experimental situations are significantly different in terms of **Performance** and **Effort**. For the effort, it could be effort linked to perception and/or to cognition processing the task.

For the global score for **Performance**, the situation "finding the route using a paper map" was rated as the one inducing the higher workload. Indeed, wavering manoeuvres were common during this session, and the driver remembered clearly these phases. Objective analysis of the performance from the experimental vehicle will indicate if there was an objective low performance in this experimental session, or if it was more a question of subjective feeling from the driver after running this session.

In terms of **Performance** as well as for **Effort**, the difference between the two situations "use of complex ergonomic mock-up" and "use of paper map" did not happen to be significant. The DALI values indicated that there was a higher global load when using the ergonomic mock-up, in comparison with the paper map, but that there was no load due to auditory perception in this second situation (indeed, there was no auditory messages when using a paper map).

In terms of **Performance** as well as for **Effort**, the difference between the two situations "human co-pilot" and "guidance system" did not happen to be significant according to the statistical analysis. The DALI values indicated there were no significant differences in terms of visual, auditory, attention, interference and temporal demand between these two situations, but a difference in terms of stress, due certainly to the fact that the driver was not familiarized with the guidance system.

The use of the on-board guidance system and the guidance by a co-pilot was rated as safe in terms of **Performance** and **Effort**, according to a safety threshold predefined.

Results both for Performance and Effort indicated that the situation where the driver has to rely on a paper map to find his route (consultation of the map while the vehicle was stationary in order to memorize the route) was an unsafe situation according to this threshold.

Considering the set of diving sub-tasks, the "itinerary following" is above the threshold of safety criteria both in terms of **Performance** and **Effort** in the situation where the driver was using the paper map. The level of **Effort** for "ability to react to dynamic road environment" was also above the threshold of safety in the situation of "using the complex ergonomic mock-up". These cases were the only ones where the threshold of safety was overreached.

## Summary of main results from PSA-TLX Compromise Effort/Performance

The computation of the compromise between Performance and Effort is justified, as it is valuable to take into account both types of answers.
Indeed, the effort induced by a task, and the assessed cost by the driver, is not necessarily correlated with the performance of the task. Classically, a driver can have the strategy to keep the same level of performance but it will be more or less costly according to the complexity of the task. Nevertheless, when considering this classical finding in ergonomics, it is referring to the objective performance and not the subjective one assessed by the driver.

### 6.2.3 Analysis of the results from BMDMW

- Variation of workload across conditions

There is a possibility to get a global score from the BMDMW questionnaire by averaging the scoring from the set of the 33 items (see Fig. 6-21).

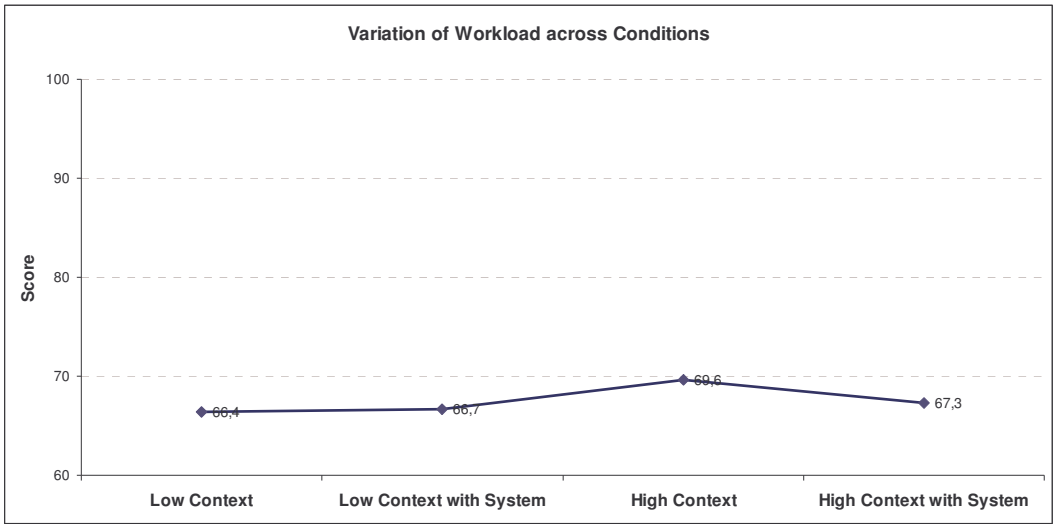



**Fig. 6-21 : BMDMW-Variation of global workload across conditions**

This score showed that the High Context session "use of the paper map" is slightly above the 3 other conditions. This global score did not seem to be very sensitive, certainly because it is the mixing items of diversified origins, the effects of some of them being annulated by the effects of others (see Table 6-14).

| *Wilcoxon test* | Conditions LC - LCS | Conditions LCS - HC | Conditions HCS - HC | Conditions LCS - HCS | Conditions LC - HC | Conditions LC - HCS |
|---|---|---|---|---|---|---|
| Global workload variation across conditions | **NS** | **NS** | **NS** | **NS** | **NS** | **NS** |

**Table 6-14 : BMDMW-Wilcoxon test used to compare global workload for the experimental sessions**

In order to compensate this problem, CRF proposed to classify the set of items in two categories: items corresponding to **positive aspects of driving** (questions 1, 4, 8, 10, 15, 16, 17, 18, 21, 25, 26, 27, 28, 29, 30, 32 and 33) and items corresponding to **negative aspects of driving** (questions 2, 3, 5 ,6, 7, 9, 11, 12, 13, 14, 19, 20, 22, 23, 24 and 31).

- Positive aspects of driving

There is an ambiguity with the item "Reach destination without paying attention to path" where the answer varied from "rarely" to "often", but there is only one destination by session so it is not logic to ask about rating the frequency of this statement (see Fig. 6-22). For the other items, the variability of the experimental sessions did not induce that much difference for each of the asked questions.

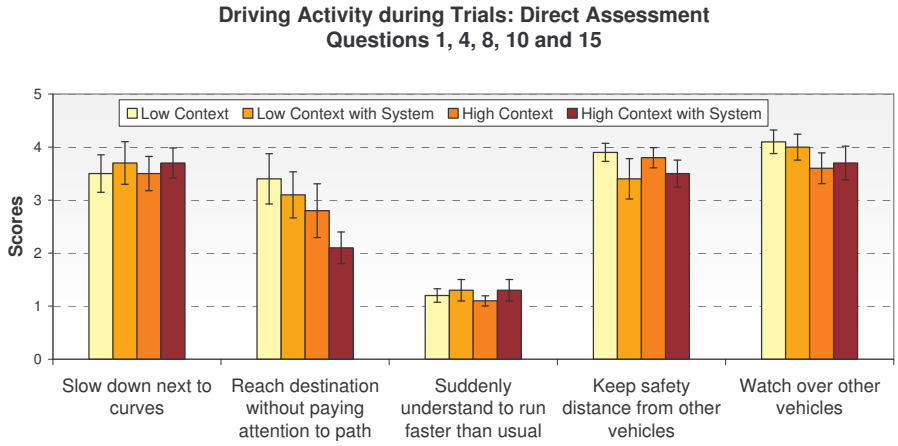



**Fig. 6-22: BMDMW-Scores got for the questions 1, 4, 8, 10 and 15**

For the item "Have no difficulty in deciding whether go straight or turn", the result is not easy to interpret. Indeed, the higher value is observed in the session "use of paper map", indicating that it is in this situation that the drivers encountered the more often (rating going from "rarely" to "often") the less difficulty. Considering the results from the two other tools and setting up of the experimental procedure, this result is the opposite of what would have be expected. One possible hypothesis would be that the rating scale was not easy to understand by the driver (see **Erreur ! Source du renvoi introuvable.**).

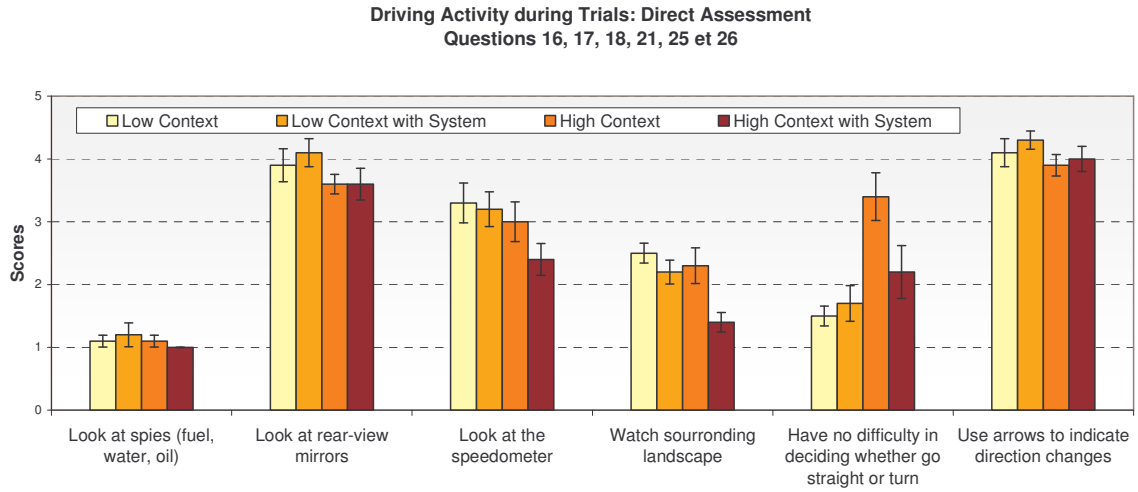



**Fig. 6-23: BMDMW-Scores got for the questions 16, 17, 18, 21, 25 and 26**

The item "suddenly understand to run slower than usual" displayed the higher value (corresponding to "often") for the session about "using paper map". Observation inside the vehicle by experimenter indicated that the drivers' attention was so deeply involved in finding his way, relying on his memorised route, and watching road environment, that he had a tendency to slow down (see Fig. 6-23).

The item "Slow down in case of low visibility (fog, rain)" has no validity in this experimental context as the sessions had been conducted in July with a good weather.

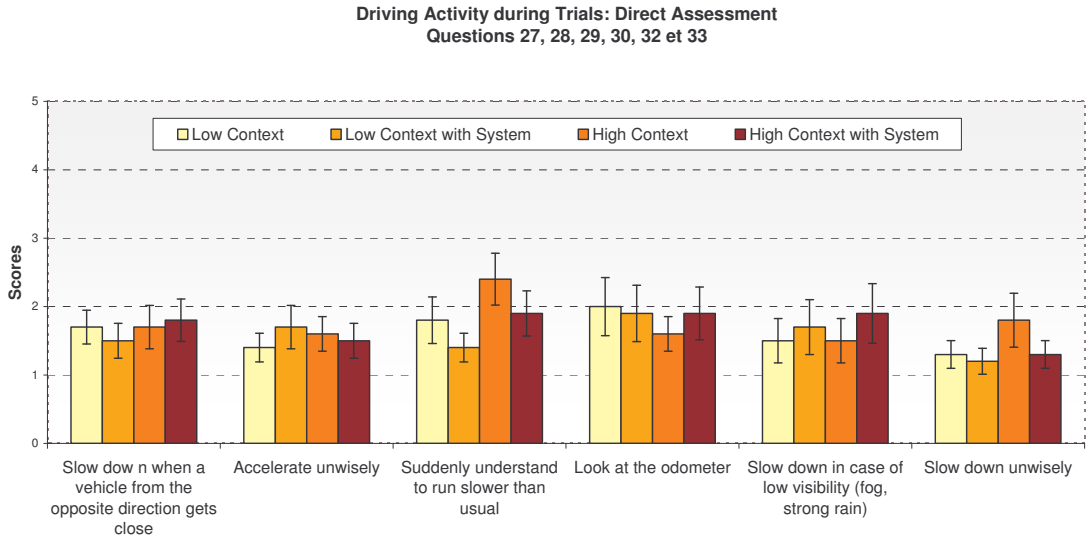



**Fig. 6-24: BMDMW-Scores got for the questions 27, 28, 29, 30, 32 and 33**

Comparison of sessions - Summary of the significant results

| Wilcoxon test | LC – HC | LC – HCS | LCS – HC | LCS – HCS | HCS – HC |
|---|---|---|---|---|---|
| Watch surrounding landscape | - | S**** | - | S* | S* |
| Have no difficulty in deciding whether go straight or turn | S*** | - | S* | - | - |
| Look at rear-view mirrors | - | - | S** | S** | - |
| Keep safety distance from other vehicles | - | S* | - | - | - |
| Use arrows to indicate direction changes | - | - | S* | - | - |
| Look at the speedometer | - | - | - | - | S* |

**Table 6-15 : BMDMW-Wilcoxon test used to compare scores of direct assessment for the experimental sessions**

Among the 17 questions linked to the positive aspects of driving, few results showed a significant difference between the Low and High conditions (see Fig. 6-24):

- Watching surrounding landscape became more difficult when the driver had to deal with different tasks as he did for the HCS condition.

- "Have no difficulty to decide whether go straight or turn" showed that using a paper map to navigate increased the number of time that the driver didn't know where to turn and to go.

- If "Look at rear mirror" was "easier" when the driver is helped by the guidance system, in the HC and HCS conditions, the global tasks were more demanding so the

driver had to change his way to drive. However, the results points out a significant difference between the LC and the LCS conditions and no significant difference between LC, HC and HCS conditions which seems to be very strange.

Even if these results fit with the different types of session, several other results are more difficult to explain:

- The questions connected with "Keep safety distance from other vehicles", "Use arrows to indicate direction changes", "Look at the speedometer" have only one significant result between two sessions which are different from one question to another.

- Negative aspects of driving

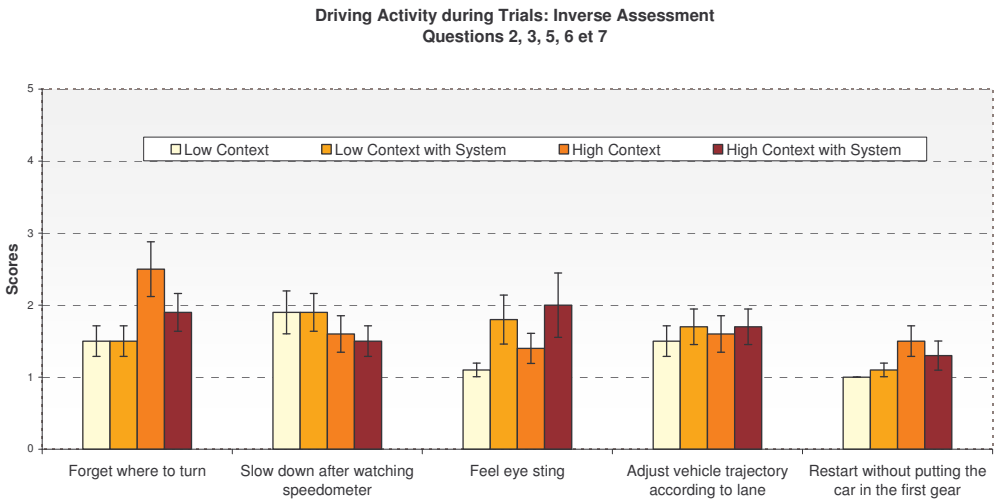



**Fig. 6-25: BMDMW-Scores got for the questions 2, 3, 5, 6 and 7**

For the item "do other things during driving", the highest values are for the situation "use of complex ergonomic mock-up". This item is valid to reveal extra activities if there are loading the driver. Indeed, the situation "using the guidance system" did not display high values (see Fig. 6-25).

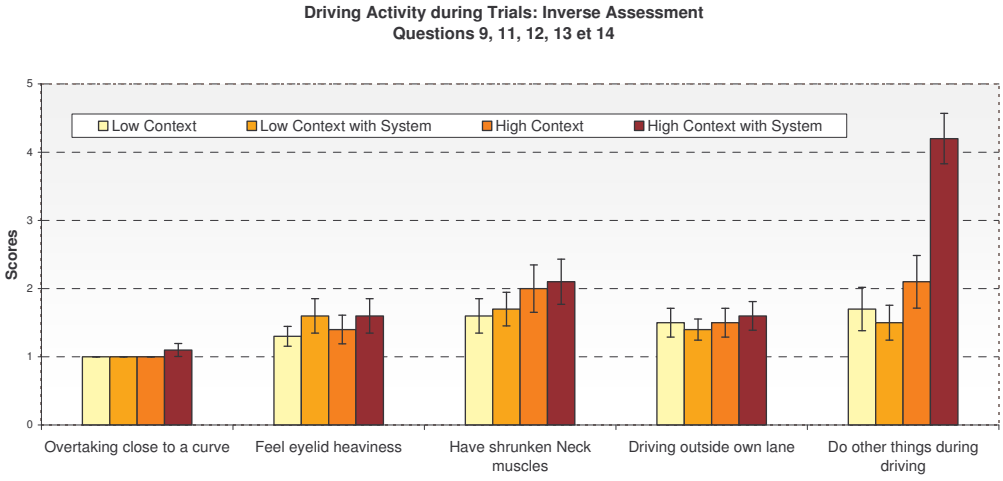



**Fig. 6-26: BMDMW-Scores got for the questions 9, 11, 12, 13 and 14Figure 53**

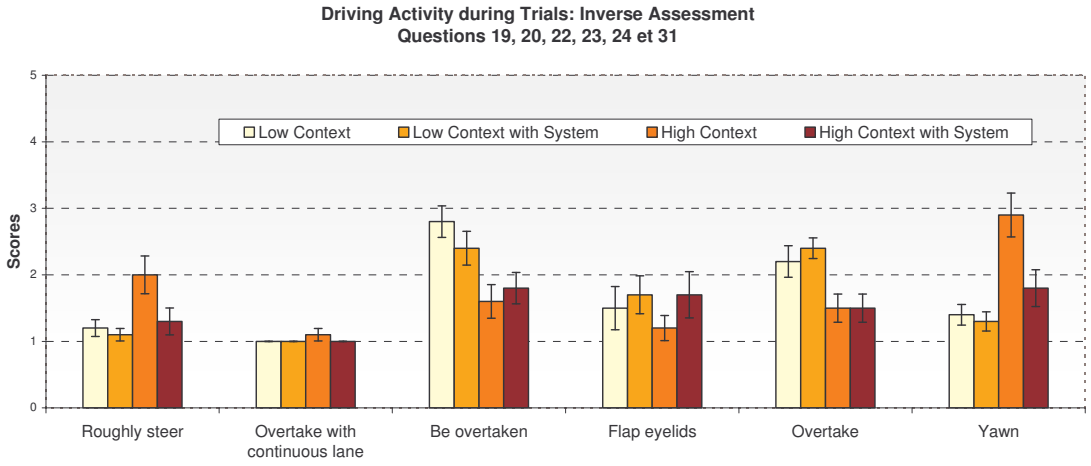



**Fig. 6-27: BMDMW-Scores got for the questions 19, 20, 22, 23, 24 and 31**

Comparison of sessions - Summary of the significant results (see Table 6-16)

| Wilcoxon test | LC – HC | LC – HCS | LCS – HC | LCS – HCS | HCS – HC |
|---|---|---|---|---|---|
| Overtake | **S*** | **S*** | **S***** | **S***** | - |
| Be overtaken | **S**** | **S**** | **S*** | - | - |
| Do other things during driving | - | **S***** | - | **S***** | **S***** |
| Roughly steer | **S*** | - | **S**** | - | - |
| Yawn | **S**** | - | **S***** | - | - |
| Restart without putting the car in the first gear | - | - | **S*** | - | - |

**S***: p<0, 05 **S****: p<0, 025 **S*****: p<0, 01 **S******: p<0, 01

**Table 6-16: BMDMW-Wilcoxon test used to compare scores of invers assessment for the experimental sessions**

Among the 16 questions linked to the negative aspects, few results indicate a significant difference:

- In the HC and HCS conditions, the driver thinks that he had more difficulties to overtake comparing to the LC and the LCS conditions. According to the results, simultaneously using a guidance system and describing information or searching the road with a paper map increase the number of time the driver thinks he has been overtaken. However, there is no significant difference between LCS and HCS conditions.

- In the HC condition, the number of time the driver has the feeling “roughly steer” and “yawn” was higher than in the other conditions; the driver had to search the route and

paid less attention to the driving task. For the HCS condition, the driver can choose to privilege the driving task with a detriment of the secondary task's management.

- The driver had a feeling of having managed a lot of things in the HCS condition.

Nevertheless, the result linked to "Restart without putting the car in the first gear" is more difficult to interpret: Why did drivers generally think that there was one significant difference only between LCS and HC conditions? It is easily to understand why in the HC condition "Restart without putting the first gear" didn't become higher than the LC condition?

## Summary of main results from BMDMW

Regarding the questions linked to the positive and negative aspects of driving, few results show a significant difference between the sessions. Globally, the results are difficult to explain and they are not always related to the experimental session complexity, or, at least, not systematically. This fact is certainly due to the diversity of the 33 items, that are covering various aspects of the driving task and the driver's state.

The BMDMW is a newly developed tool which would deserve to be further investigated.
The easiness of the process of data gathering and data processing is valuable and should be kept.
The type of items should be classified under categories in relation to their meaning in terms of drivers' workload, driver's actions, driver's state… Some items such as "Yawn" could be withdrawn: in this example, there is an uncertainty if yawning is due to the drivers' fatigue (and if the fatigue is really due to the experiment or previous activities before the experiment), to the driver's boredom or just by chance during one or the other session.

### 6.2.4 Synthesis of final evaluation of questionnaires

No differences are observed between BMDMW, DALI and PSA-TLX regarding:

- Score's choice : "easy"
- Questions understanding :" well"
- Concepts understanding :" well"
- Written explanations understanding: "easy"
- Evolution of effort (learning effect) : "easier" or "equal complexity"
- Exhaustivity
- Effort required to fill in both questionnaires: "moderate"
- Explanations given by the experimenter: "necessary"

Significant differences between BMDMW, DALI and PSA-TLX regarding:

- Questionnaire's presentation: better presentation for the DALI (95% vs 70% for the PSA-TLX and 60% fir the BMDMW)
- Duration of questionnaire filling: quick filling in of the BMDMW and the DALI questionnaires (90% and 85% vs 50% for the PSA-TLX questionnaire)
- Explanations given by the experimenter: sufficient for BMDMW and DALI questionnaires (100% for the BMDMW and the DALI vs 90% for the PSA-TLX)
- Level of effort required: low effort and moderate effort for the BMDMW and the DALI (100% and 95% vs 60% for the PSA-TLX)
- Driver'state at the end of the experiment: active when the experiment is with DALI + PSA-TLX (30% vs 80% with the DALI + BMDMW)

### *6.3 Discussion on INRETS results*

Regarding the results from the final questionnaire, there is no much significant difference between the 3 tested tools (score's choice, questions understanding, concepts understanding, written explanations understanding, evolution of effort along trials, exhaustivity, and explanations given by the experimenter). However, according to the results from the final questionnaire, the presentation of BMDMW and PSA-TLX questionnaires could be improved.

The **DALI** tool allowed to show significant differences between the experimental sessions in terms of perceptive, cognitive, stress, temporal demand and interference induced by the driving task.
The DALI questionnaire has the advantage to allow the identification of the origins of the workload for the driver, leading then to correct the situation at this identified level (e.g. interference and visual load indicate that an in-vehicle system will have a visual demanding visual display). Some possible improvements have to be made:

- adding factors linked to specific aspect of the driving task would be useful to evaluate impact of ADAS (e.g. level of stress to keep distance with the vehicle ahead, in the case of a system having an impact on this specificity of the driving task)

The **PSA-TLX** tool allowed showing significant differences between the experimental sessions in terms of evaluated performance and effort by the driver.
The PSA-TLX questionnaire allows evaluating the level of subjective performance while using a system (possibility to compare with the objective performance and to identify in which way the system can modify the risk taking, for instance). This approach has to be compared with a reference situation with no system, as evaluated performance can be also twisted by drivers that can overestimate their driving performance. With the PSA-TLX, it is possible to evaluate the level of effort felt by drivers for various identified driving sub-tasks.

However, in our experiment, the item "Ability to react to safety and status indicators" did not seem to be necessary: in fact, as it had been explained before, the vehicle used doesn't belong to the driver. So, in driver's mind, the experimenter must check the vehicle before beginning the experiment. The items "Control of the lateral motion" and "Control of the longitudinal motion" were difficult to evaluate by the drivers: in fact, generally, drivers had difficulties to analyse these two types of control because they are highly automatic for them. Indeed, subjective evaluation concerns only to conscious processes.
Furthermore, there are too many questions, sometimes complex, to be answered by the driver in a row: the problem in this case is that the driver has a tendency to forget the characteristics of his driving task as he has to concentrate on the meaning of the questionnaire.
Furthermore, in the template of analysis proposed by PSA at this stage, there are also too many proposed tests and graphs consuming a lot of time for the experimenter, but this can be easily improved.

Therefore, some possible improvements could be made in order to produce a more synthetic final report:

- making the questionnaire lighter by withdrawing some questions
- making the process of data gathering and data analysis much more lighter.

The **BMDMW** tool did not allow bringing systematic evaluation of driving workload according to the experimental session. Even if it is easy to process at all the stages of the data

gathering and computation, a categorisation of items in relation to their meaning for the driving task could possibly improve the questionnaire.

The figure below synthetises what the 3 tools allows to evaluate:

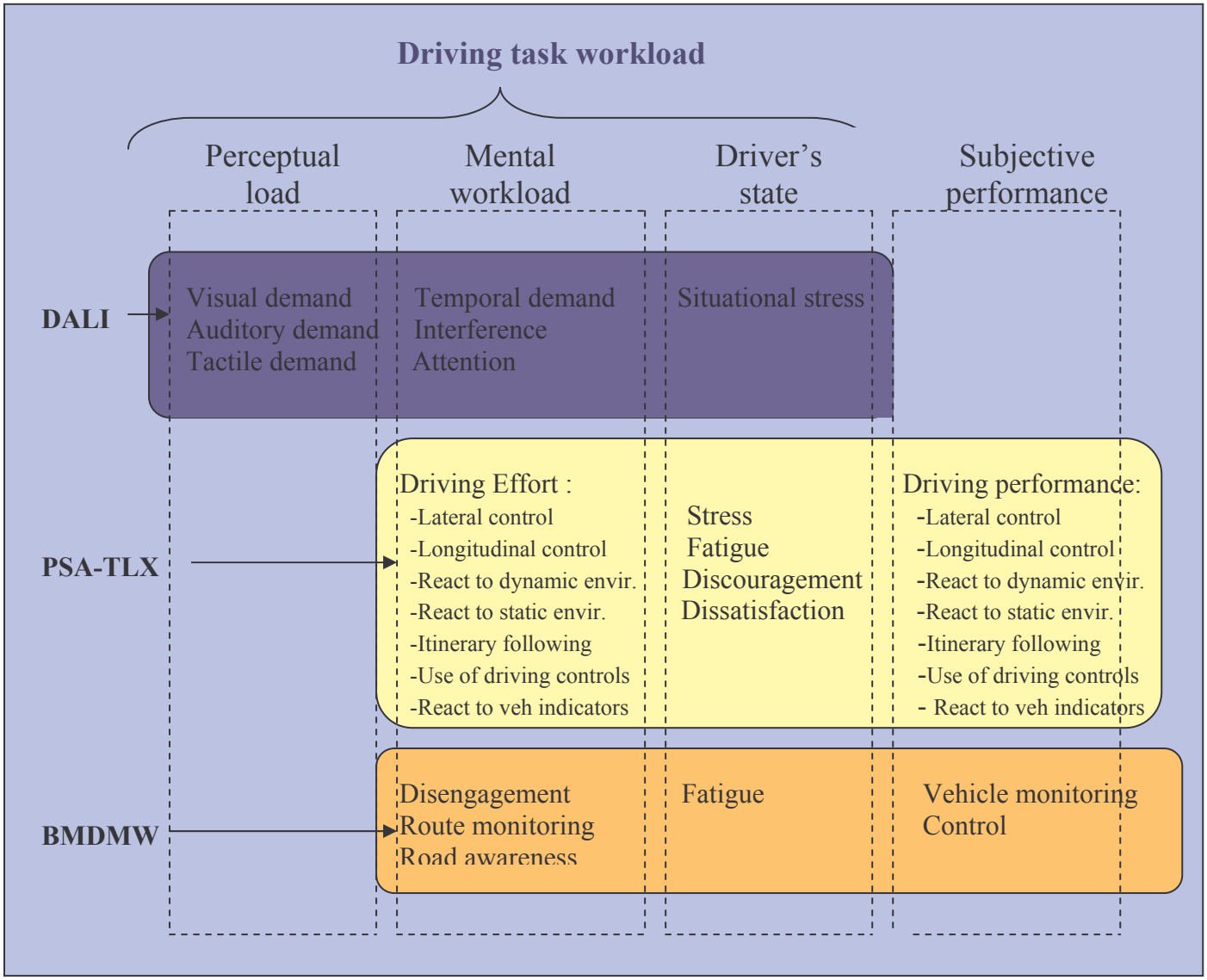



**Fig. 6-28: BMDMW-Scores got for the questions 19, 20, 22, 23, 24 and 31**

# 7 Experiment 3 (PSA)

The aim of this experiment is to assess the sensitivity of two subjective methods dedicated to evaluate either mental workload and/or driving workload when using an Adaptive Cruise Control on open road.

Two different tools are evaluated :

- **The PSA-TLX (PSA-Task Load index)** (see section 2.4.3)
- **The DALI (Driving Activity Load Index)** (see section 2.4.2).

BMDMW was not tested as not available when starting the road tests, and this date couldn't be postponed.

## *7.1 Method*

### 7.1.1 Participants

12 male drivers have participated to the study. Their characteristics are:

- 23 to 40 years old
- Driving experience of 10 000 Km per year and driving licence of 5 years at least
- At ease with driving a *big model* of vehicle (as the C5 test vehicle)
- At ease with using a automatic gearbox
- At ease with driving on high speed road
- Never having driven with the ACC nor the distance warning system
- Knowledge and/or experience with the conventional cruise control but not used to it.

### 7.1.2 Apparatus

The test vehicle is a Citroën C5 with automatic gearbox, equipped with an adaptive cruise control (ACC) (see Fig. 7-1 and Table 7-1)

**Fig. 7-1 : Test-vehicle**

| Data | Devices |
|---|---|
| Video recording | 4 cameras (2 for driver's eye, 1 for the scene ahead and 1 directed to pedals |
| | Image mixer |
| | video tape recorder |
| Supply | Transformer 12V - 220V |
| | Tranformer 220V – 12V |
| Vehicle data collection | Laptop |
| | Software |
| | Monitor for supervision |

**Table 7-1 : In-vehicle devices**

### 7.1.3 Data analysis

- DALI measures

See section 2.4.2

- PSA-TLX measures

See section 2.4.3

- Workload questionnaires' evaluation: final questionnaire

See section 5.1.4

### 7.1.4 Inferential statistical measures

To compare the conditions, two statistical measures (common to all methods and all partners) are used :

- Wilcoxon test
- Rank correlation of Spearman.

These tests have been chosen as non parametric test suitable to low total number of scores.

### 7.1.5 Comparison of conditions

Sensitivity of the PSA-TLX and the DALI is evaluated through the comparison of these conditions :

- condition 1 : driving without the system
- condition 3 : driving with the system after the familiarisation with the system
- condition 4 : driving without the system after driving with the system.

No questionnaire has been administrated in condition 2 after familiarisation with the system (see Table 7-2)

| Comparison of conditions | Driving without ACC (1) | Driving with ACC – familiarisation (2) | Driving with ACC (3) | Driving without ACC (4) |
|---|---|---|---|---|
| Driving without ACC (1) | | | | |
| Driving with ACC – familiarisation (2) | - | | | |
| Driving with ACC (3) | X | - | | |
| Driving without ACC (4) | X | - | X | |

**Table 7-2 : Comparison of conditions**

### 7.1.6 Tasks/Systems (e.g. IVIS tasks or ADAS systems, both…)

- **Driving task without the system**

All trials are carried out in daytime, on a high speed open road and motorway in order to have participants driving at speed range compatible with the system functioning limits.
Time of trials has been chosen to avoid high density traffic and traffic jam.

In total, the route is composed by :
300 km (150xA/R) in 3 or 4 lanes highways (A10) and
30 km (15xA/R) of 2- lanes national road to reach highways (high traffic density and speed limitation to 90km/h or 110 km/h) , but this part has not been taken into account in the four test-conditions.

Instructions given to drivers were to drive the more naturally with their spontaneous speed. This should avoid generating an additive source of workload due to an artificial speed management.
Speed limitation in highways is 130km/h or 110km/h depending the portions (see Fig. 7-2)

**Fig. 7-2 : Itinerary**

- **Driving task with the system**

System's description :

The evaluated system is an assistance system whose purpose is to help to control and homogenise traffic (if all vehicles are equipped with it) and reduce potential risk of accidents.
The system provides an assistance for longitudinal control: speed and time headway management. The system is equipped with a radar gathering the relative speed and the relative distance to the front vehicle and transfers these data to a calculator.
The system works between 60km/h and 160km/h and does not detect fix obstacles (e.g. wall).
Two functions are integrated :

- Automatic mode: automatic management of vehicle speed (function of adaptive cruise control) within two parameters selected by the drivers: target time headway and target speed.
- Informative mode: function of distance warning when effective time headway is lower than target time headway selected by the driver, when system is activated and the automatic mode is not activated

Automatic control mode of speed : ACC

The system achieved the longitudinal control of the vehicle : speed management and distance management if a preceding vehicle is detected, depending on :

- the target speed (between 60km/h and 160km/h) and target time headway (between 1s and 2,5 s., resolution 0,1s.) selected by the driver and
- data collected by the radar : relative distance and relative speed from the front vehicle
- current speed of the host vehicle.

The system can brake (brake light on), decelerate and accelerate according to the parameters selected by the driver. When no target is detected, the ACC function is to reach the target speed as a conventional cruise control.

Information displayed to the driver is (see Fig. 7-3):

- Time headway (target and effective) in red or green depending on the comparison between the effective time headway and the target time headway,
- Target speed,
- State of the system (OFF, ON, over…),
- Target detected or not : a drown "full car" means a detected target, while "an empty car" means that no target is detected (despite the ACC function activated)

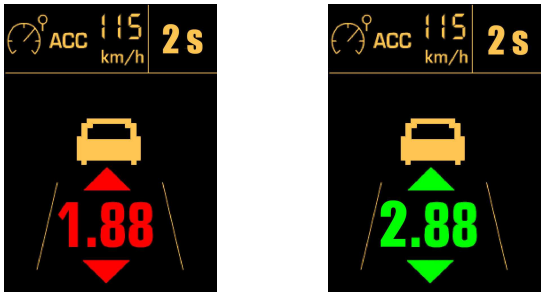



**Fig. 7-3 : ACC system HMI**

Whatever he wants, the driver can resume the control of the vehicle:

- deactivate the system by braking and the driver must switch on by manual control to activate the function
- inhibit the system by accelerating (temporarily, during the action on accelerator), and as soon as the accelerator is released by the driver, the system resumes.

Informative mode (distance warning system): DWS – Distance Warning System :

This mode is activated by default when the automatic ACC function is deactivated by the driver with manual control or braking (without deactivating the system).
The system warns the driver if the current time headway to the preceding vehicle is inferior to the target headway set by the driver as his safety headway.
The warning is a vibration in the accelerator pedal to indicate that the time headway is unsafe, and the warning is visual, as the time headway is displayed in red.
Information displayed to the driver is the same as for the automatic mode, without ACC information (target speed, ACC, ACC states):

- time headway (target and effective),
- a full coloured vehicle.

Drivers are free to select and change the target time headway and the target speed, and to deactivate the system or one of both functions, whenever they want (see Fig. 7-4)

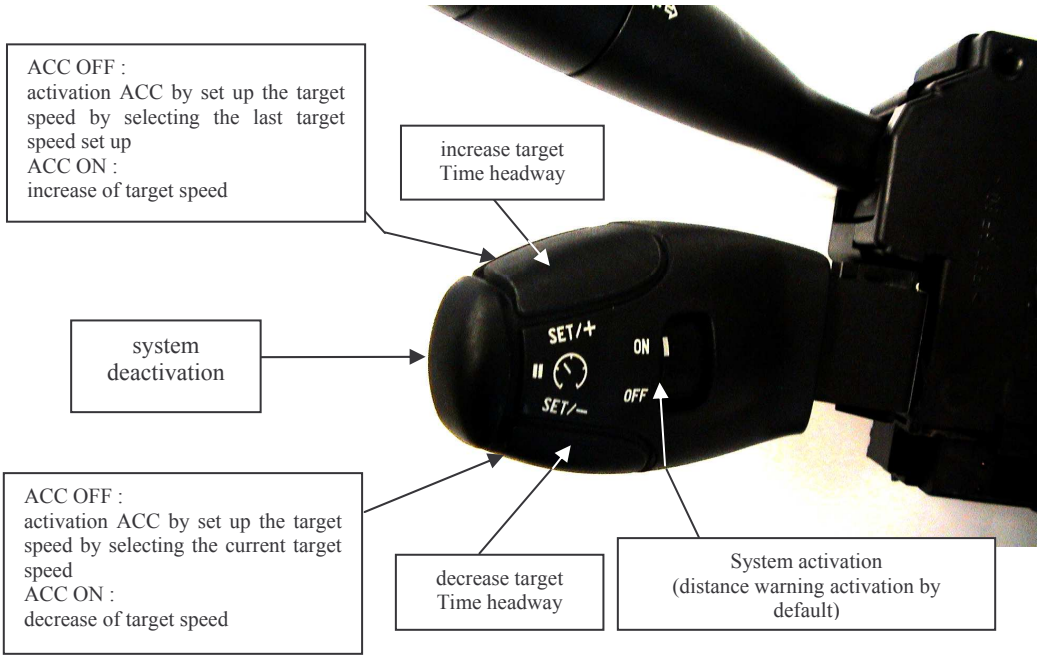



**Fig. 7-4 : Manual controls of ACC system**

Driving with ACC :

The function performs a part of the driver's vehicle control: the longitudinal control by management of the vehicle speed.
Considering that the duration of system use (conditions 2 and 3) is around 90 minutes, the driver's task is to learn to use the system e.g. discover and identify:

- The limits of the system : the situations the system performs a wrong diagnosis of the situation (limits of the system) : e.g. in curves (the system may detect a wrong target present in other lane and brake or may loose the target and accelerate to reach the target speed)
- Normal actions of the system perceived as abnormal working: situations the system actions are not appropriate to the drivers' intention: e.g. before overtaking.

Driver has to supervise the actions of the system and diagnose if they are appropriate to the situations. So the situation awareness may increase.

The other key thing is that the driver must pay attention to the rear scene as the system's deceleration and braking are only determined by vehicle ahead, rear vehicles are not taken into account by the system.

When no target is detected, the ACC function is a conventional cruise control, the speed is controlled to reach the target speed and to not exceed it. Driving task deals with lateral control and taking into account road situation (road users and infrastructure/ traffic signs).

Driving with the distance warning system :

The function is to warn the driver when the current time headway is lower than the chosen target time headway. The function is informative, so the driving task is not modified: control of the vehicle (longitudinal and lateral control) is achieved by the driver.

### 7.1.7 Design and Procedure

The experimentation has 4 main conditions (see Table 7-3):

- Driving without the system (40 minutes) : 2 sessions – conditions 1 (before the system use) and 4 (after the system use) -
- Driving with the system (90 minutes) : 2 sessions – conditions 2 (familiarisation) and 3 -

Workload evaluation questionnaires – PSA-TLX and DALI – are filled after three of the four conditions : 1, 3, 4, as the condition 2 deals with the familiarisation with the system.
Workload is estimated by novise drivers with ACC : after 90 minutes driving with the system, drivers are expected to be fully unacquainted with the system.
One final questionnaire is administrated at the end of the experiment, to collect drivers' point of view on the tools PSA-TLX and DALI.

At the beginning of the experiment, drivers are not aware of the methodological issue and workload methods evaluation, to avoid any bias. Drivers are informed that they will fill two independent questionnaires dealing with different concept of workload. They are not informed of the final questionnaire designed to evaluate the PSA-TLX and DALI.

| | Driving | | | Questionnaire | | | |
|---|---|---|---|---|---|---|---|
| **Test Condition** | **Without system** | **Training with the system** | **With system** | **PSA-TLX timing** | **DALI timing** | **Final questionnaire** | **Total duration (min.)** |
| (1) | 20 | | | 20 | 10 | | 50 |
| (2) | | 30 | | | | | 30 |
| (3) | | | 60 | 5 | 5 | | 70 |
| (4) | 20 | | | 5 | 5 | 10 | 40 |
| Total | 40 | 90 | | 30 | 20 | 10 | 190 |

**Table 7-3 : Design procedure (minutes)**

## *7.2 Results*

### 7.2.1 PSA-TLX

- **Driving performance**

Global performance

Average global scores show low disruption and good performance without the system, in particular in the second session of driving after using the system (condition 4). When drivers have the system at their disposal (condition 3), mean global scores show that disruption increases, but the global performance of driving is acceptable, as only few minor errors of driving are reported (see Fig. 7-5).

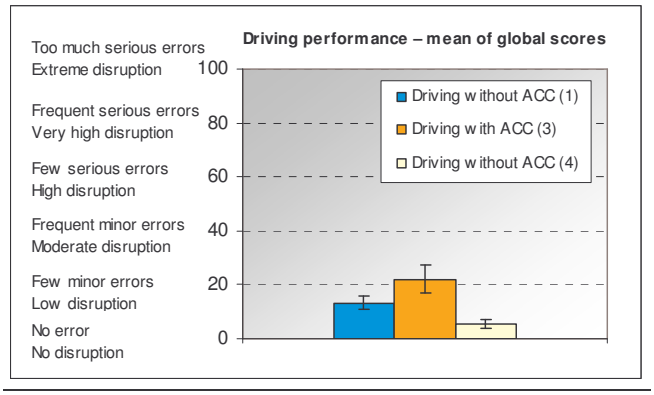



**Fig. 7-5 : PSA-TLX - Driving performance – mean of global scores**

Individual analysis of global performance of driving indicates same tendency as the group trend on the comparison between driving with the system (condition 3) and driving without the system (condition 1 and 4), and shows that the increased disruption and degradation of global performance when using the system concerns (see Fig. 7-6):

- less than a half of the drivers (5/12) when comparing driving with the system (condition 3) with the first session of driving without the system (condition 1)
- two third of the drivers (8/12) when comparing driving with the system (condition 3) with the second session of driving without the system (condition 4), after driving with it.

Among these drivers who mention a degradation of the global performance, one driver (number 6) reports high disruption and reaches the rejection threshold when using the system, showing that the system use has consequences for driving task.
For two drivers, performance is acceptable despite the increase of disruption.

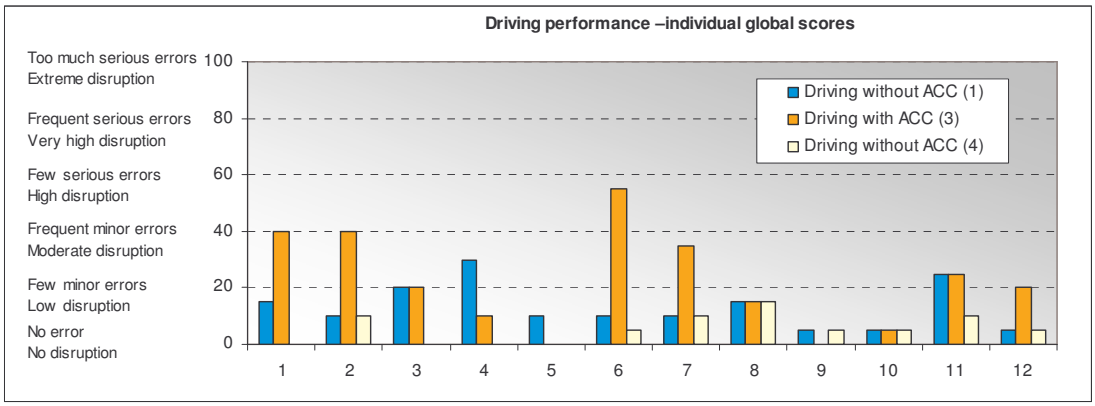



**Fig. 7-6 : PSA-TLX - Driving performance –individual global scores**

The difference can be due to:

- an increasing time exposure to the system
- driving conditions changes (more dense traffic before using the system)
- training to the vehicle
- training to the questionnaire

Performance in driving sub-task

Group trend of scores gathered for each sub-tasks of driving - lateral control, longitudinal control, reactivity to dynamic road environment, reactivity to static road environment, itinerary following, use of driving controls and reactivity to in-vehicle safety indicators - shows a slight degradation of the performance in all different sub-tasks of driving, when driving with the system (condition 3) comparing to driving without the system (conditions 1 and 4), (see Fig. 7-7 ).

Difference is significant between driving session with the system (condition 3) and the last driving session without any assistance (condition 4), as demonstrates the rank correlation of Spearman ($r = 0,77$), see and Table 7-4.

Wilcoxon test shows significant differences between (see Table 7-5):

- first driving without the system (condition 1) and driving with the system (condition 3) in performance in taking into account infrastructure and road signs (reactivity to static road environment), that is more disrupted when using the system, but mean performance is acceptable as disruption is low.
- Second driving without the system (condition 4) and driving with the system (condition 3), in performance in taking into account road users (reactivity to dynamic road environment), in using driving controls and in reacting to in-vehicle safety indicators that is slightly degraded, in average, with the assistance; but mean performance is acceptable as disruption is low.

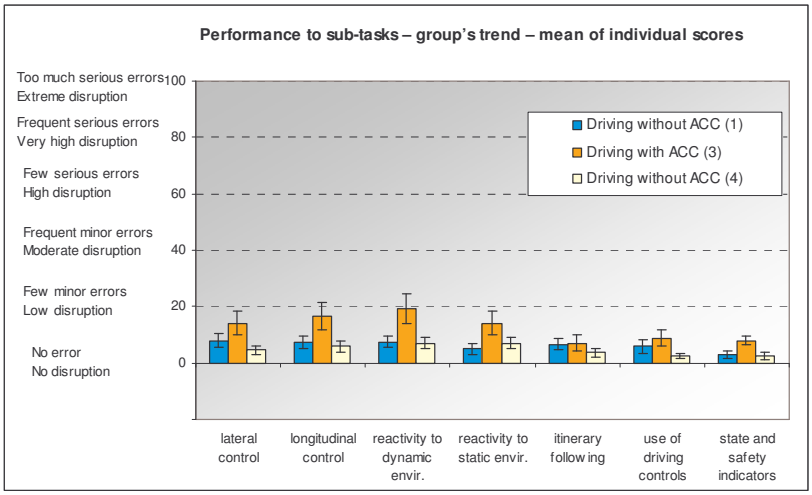



**Fig. 7-7 : PSA-TLX - Performance to sub-tasks – group's trend – mean of individual scores**

| Rank correlation test Spearman (.05)<br>Global performance | R corrected |
|---|---|
| Driving without ACC (1) / Driving with ACC (3) | 0,54 |
| Driving with ACC (3) / Driving without ACC (4) | **0,77** |
| Driving without ACC (1) / Driving without ACC (4) | 0,38 |

**Table 7-4 : PSA-TLX - Performance to sub-tasks - Spearman Ranks' Correlation test**

When comparing rank correlation of mean scores of performance to sub-tasks, the only significant difference is between the condition of driving with system (3) and the last condition of driving without system (4).
The performance is better when driving without the system, just after driving with the system. However, the performance to different sub-tasks is good in both conditions, despite it is slightly degraded when driving with the system (4).

| Wilcoxon on raw scores | Lateral control | Longitudinal control | Reactivity to dynamic environment | Reactivity to static environment | Itinerary following | Use of controls and equipment | Use of safety indicators |
|---|---|---|---|---|---|---|---|
| Driving without ACC (1) / Driving with ACC (3) | NS | *NC* | NS | **0,01** | NS | NS | NS |
| Driving with ACC (3)/ Driving without ACC (4) | NS | NS | **0,03** | NS | *NC* | **0,01** | **0,005** |
| Driving without ACC (1) / Driving without ACC (4) | NS | NS | NS | *NC* | *NC* | *NC* | *NC* |
| NS : no significance; NC : no calculation ; .01 : probability associated to calculated T | | | | | | | |

**Table 7-5 : PSA-TLX - Performance to sub-tasks – group's trend – Test of Wilcoxon**

For most of the sub-tasks, average performance remains good without the system and is acceptable with it : no important error is reported, only errors without safety importance become more frequent when the driver is assisted by the system ("*adaptation to system behaviour, changing way to drive*", "*no error, but lane changing less anticipated*", "*lost of vehicle control*, *a time to adapt is required*" as drivers comments).

When drivers use the system (3), in comparison to the preceding condition where they get trained with the vehicle without the system (1), they estimate that their performance decreases in taking into account other road users (reacting to dynamic environment).

Once used with the vehicle, drivers estimate that their driving performance without the system (4) is better than driving with the system (3) for the following driving sub-tasks:

- taking into account other road users (reacting to dynamic environment)
- using driving control
- reacting to status and safety vehicle indicators.

There is no difference for all sub-tasks between the two conditions of driving without the system:

- In the first condition (1), the disruption of performance is mainly due to training to vehicle dynamics (e.g. "*time to adapt to automatic gearbox*", "*unknown vehicle*" as drivers' comments)

In the second condition of driving without ACC (4), the disruption is no more related to training to drive the vehicle but maybe more to fatigue (3rd hour of driving).

➢

➢ *Performance in lateral control*

For 5/12 of drivers, performance in lateral control is degraded when using the system (see Fig. 7-8).

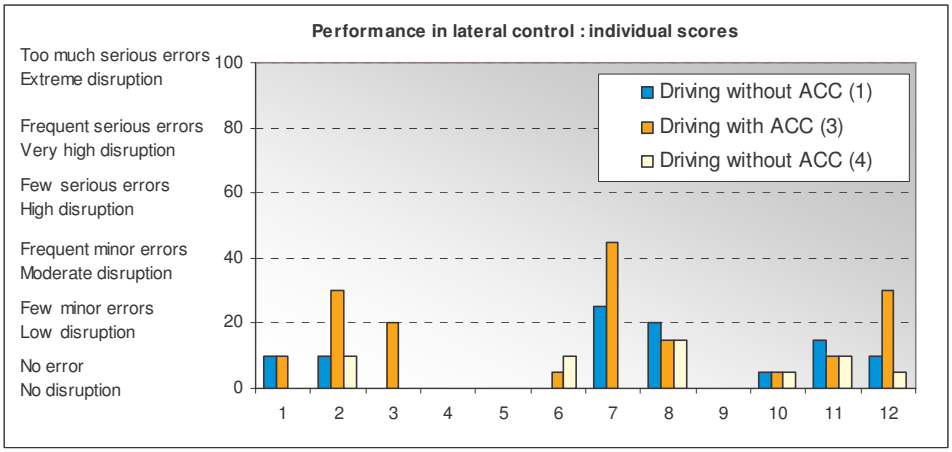



**Fig. 7-8 : PSA-TLX -Performance in lateral control – individual scores**

Indeed, this degradation of performance is more serious for one of them, who reports frequent minor errors and moderate disruption, however his score is lower than the rejection threshold showing that using the system has no major consequence for driving safety.

Among these drivers, performance in lateral control is acceptable or optimal with or without the system.

For the other 7 drivers, no difference or very little difference is found between both situations.

When comparing both conditions of driving without the system (conditions 1 and 4), few differences are significant: performance in lateral control is optimal or acceptable (low disruption).

➢ *Performance in longitudinal control*

NB : the system provides to the driver an assistance in longitudinal control, so the longitudinal control evaluated in condition with the system, results from the collaboration between the system and the driver.

For around the half of drivers, performance in longitudinal control is degraded when using the system comparing to driving without system (see Fig. 7-9)

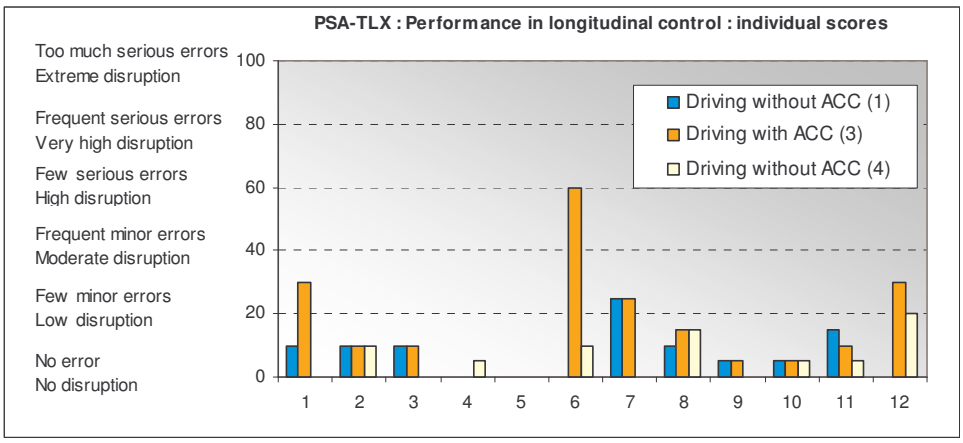



**Fig. 7-9 : PSA-TLX -Performance in longitudinal control – individual scores**

For most of the drivers, performance is acceptable or optimum with or without the system use and little difference appears between both conditions.
However one driver (number 6) is an exception : he mentions a serious disruption in longitudinal control with the system and the performance estimated is really deteriorated and unsafe (as the rejection threshold is exceeded).
This driver reports few serious errors in longitudinal control when using the system while his performance without the system is optimal. Errors deal with braking or deceleration of the system judged by the driver as not suitable with driver's intentions or unnecessary (e.g. overtaking or tolerance of short time headway to avoid braking in highways) even if system actions are normal (no bug).

When comparing both conditions of driving without the system (conditions 1 and 4), few differences are significant: performance in lateral control is optimal or acceptable (low disruption).

- *Performance in taking into account road users (reactivity to dynamic road environment)*

For around the half of drivers, performance in taking into account other road users is degraded when using the system comparing to driving without system (see Fig. 7-10)

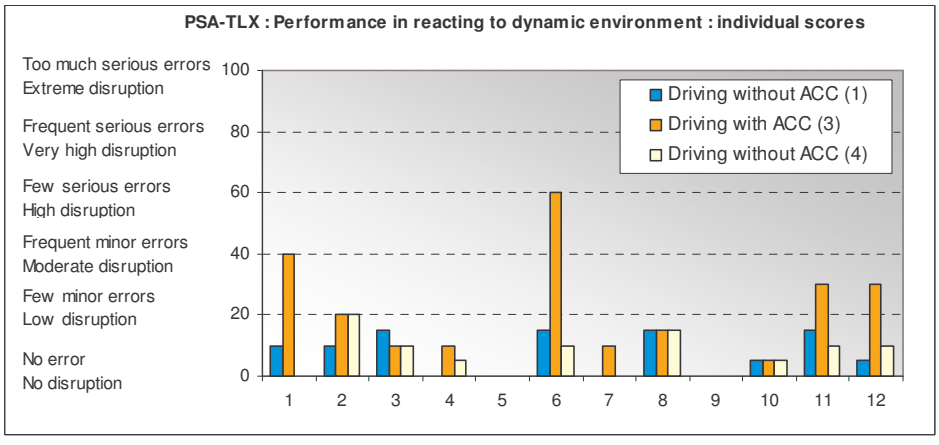



**Fig. 7-10 : PSA-TLX - Performance in reacting to dynamic environment – individual scores**

However, for most of the drivers, performance is acceptable or optimal with or without the system use and little difference appears between both conditions.

However, one driver (number 6) is an exception: he mentions a serious disruption in taking into account road users, with the system and the performance estimated is really deteriorated and unsafe (as the rejection threshold is exceeded). This driver reports few serious errors when using the system while his performance without the system is good (low disruption).
When driving with the system, driver expresses the need of a higher anticipation because of system actions that may not be suitable with his intention.

For the other half, no difference or low difference is found between driving with the system and driving with no-system.
When comparing both conditions of driving without the system (conditions 1 and 4), few differences are observed: performance in taking into account other road users is optimal or acceptable (low disruption).

- *Performance in taking into account infrastructure and road signs (reactivity to static road environment)*

For 7/12 drivers, performance in taking into account infrastructure and road signs is degraded when using the system comparing to driving without system (see Fig. 7-11)

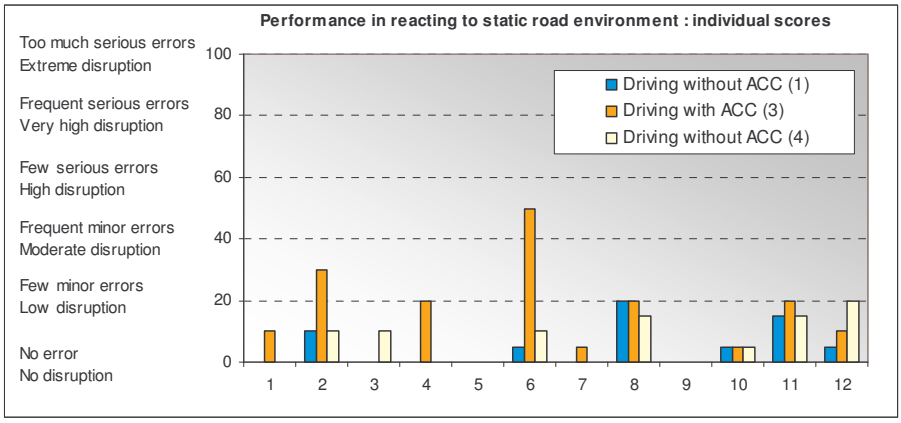



**Fig. 7-11 : PSA-TLX - Performance in reacting to static environment – individual scores**

For most of the drivers, performance is acceptable or optimal with or without the system use and little difference appears between both conditions.
However, one driver (number 6) is an exception: he mentions a moderate to high disruption in taking into account infrastructure and traffic signs, with the system and the performance estimated is deteriorated, but the rejection threshold is not reached.
This driver reports minor frequent errors when using the system while his performance without the system is good (low disruption).
For the other drivers, performance in taking into account infrastructure and road signs is acceptable and/or optimal with and without using the system.
When comparing both conditions of driving without the system (conditions 1 and 4), few differences are observed: performance in taking into account other road users is optimal or acceptable (low disruption).

➢ *Performance in itinerary following*

Globally, performance in itinerary following is acceptable or optimal in all conditions, despite small differences and variation between drivers according the different test conditions.
The majority of the drivers mention low disruption or nil disruption when driving with the system or without any assistance.
For most drivers, the itinerary is not difficult as it is in highways or known by drivers, and as experimenter guided the driver. This sub-task present low demands.

➢ *Performance in using driving controls*

Globally, performance in using driving controls is acceptable or optimal in all conditions, despite small differences and variation between drivers according the different test conditions.
The majority of the drivers mention low disruption or nil disruption when driving with the system or without the system.
When driving with the system, the use of system's controls presents few minor errors as there is differences and confusion with controls of the conventional cruise control known by some drivers.
One driver (number) reports moderate disruption as he was really influenced by controls of the conventional cruise control.

➢ *Performance in reacting to in-vehicle state and safety indicators*

Globally, for all drivers the performance in reacting to in-vehicle indicators is acceptable or very good in all conditions. No driver reports moderate disruption in all conditions.
When using the system, some drivers mention being a little bit disturbed by auditory warnings when system is inhibited, but the disruption is minor.

- Management of performance

The composition of disruption doesn't change in fundamentals (see Fig. 7-12).
- When drivers have the system at their disposal,
- When driving without the system before and after using the system.

Main changes consist in the disruption of longitudinal control task and of ability to take into account the road situation as infrastructure and road signs: the disruption increases when driving with the system. Nevertheless, observed differences are not statistically significant.

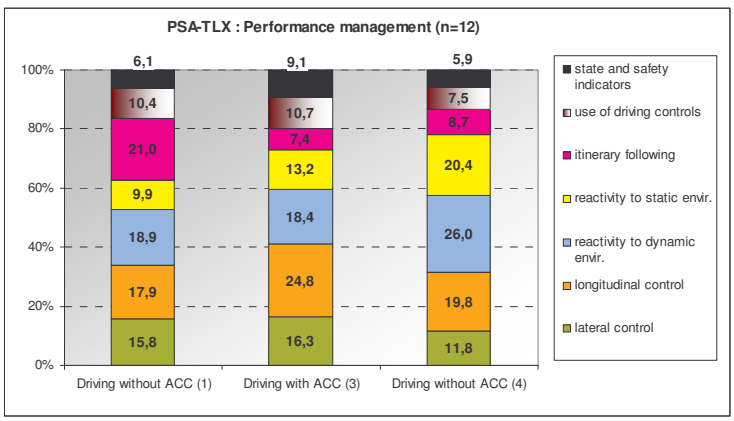



**Fig. 7-12 : PSA-TLX -Performance management**

- **Driving effort**

Global effort

Mean global effort is low when driving without the system and when driving with it: mean effort is lightly higher when driving with the system (see Fig. 7-13).

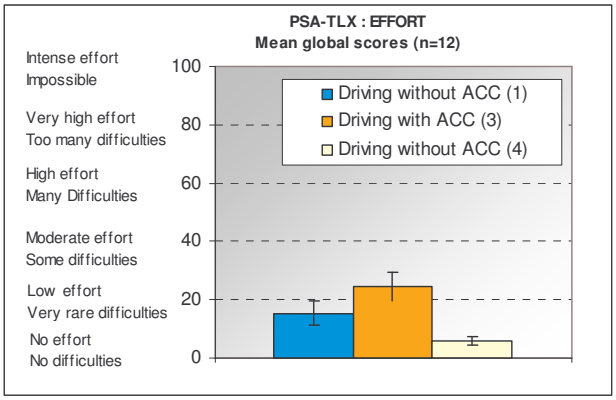



**Fig. 7-13 : PSA-TLX -Driving effort - Mean of global scores**

When driving with the system, 5 drivers consider their effort moderate, for the other, effort is nil or very low (see Fig. 7-14). The main reason given by drivers to explain the effort allocated to driving, is the modification of their driving behaviour to cope with the system behaviour. This modification concerns situations when approaching a front vehicle before overtaking, and system speed control (deceleration and braking).

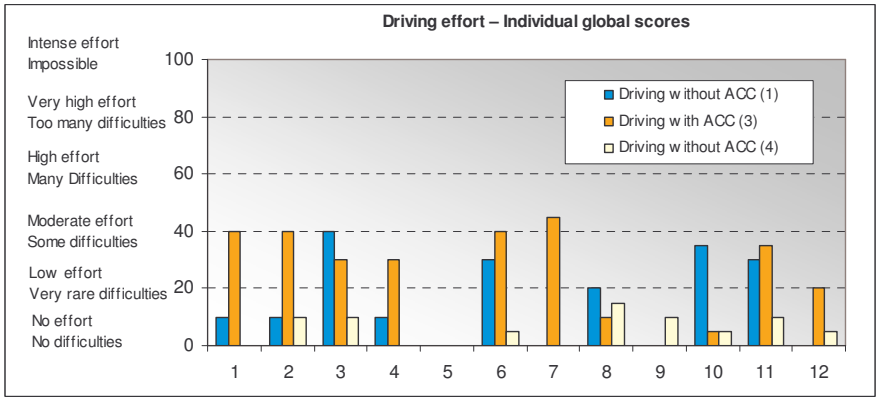



**Fig. 7-14 : PSA-TLX -Driving effort – Individual global scores**

When driving without the system after having used the system, drivers mentioned low effort or no effort. This can be explained by:

- adaptation to test situation
- adaptation to vehicle (automatic transmission, comfort, etc.)
- lighter traffic in the first test condition

There is a noticeable difference between both conditions of driving without the system (conditions 1 and 4): mean global effort of driving is lower the second time without the system (after using the system) than the first time without the system (before using the

system). This difference is explained by difference of traffic density, being heavier in the first session.

- Effort for each sub-tasks of driving

Group trend shows global effort is low in average, for all sub-tasks of driving and all conditions, except for reacting to dynamic road environment (interactions with other vehicles) where effort allocated is low to moderate when driving with the system (see Fig. 7-15)

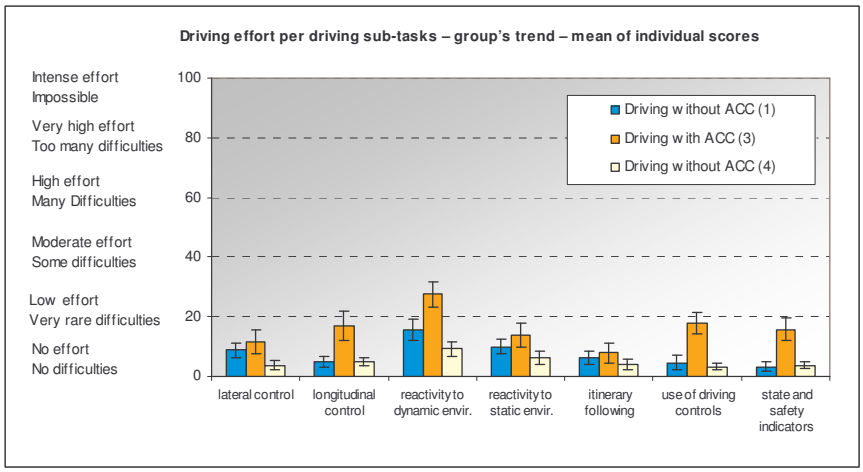



**Fig. 7-15 : PSA-TLX -Driving effort per driving sub-tasks – group's trend – mean of individual scores**

Rank correlation of Spearman indicates no difference between driving with the system (condition 3) and driving without the system (conditions 1 and 4), while shows a significant difference between both session of driving with no-assistance (see Table 7-6).

| Rank correlation test Spearman (.05)<br>Raw scores for performance | R<br>corrected |
|---|---|
| Driving without ACC (1) / Driving with ACC (3) | 0 |
| Driving with ACC (3) / Driving without ACC (4) | 0,18 |
| Driving without ACC (1) / Driving without ACC (4) | **0,76** |

**Table 7-6 : PSA-TLX -Effort to sub tasks- Spearman ranks' Correlation test– raw score**

Wilcoxon test demonstrates higher mean effort for using driving controls and for reacting to in-vehicle indicators, when using the system (condition 3) comparing to both conditions of driving with no-assistance (conditions 1 and 4). There is no difference between both situations of driving without the system, except in effort for the lateral control, sub-task for which effort supplied is little higher when getting used with the system (condition 1), see Table 7-7.

Few differences are observed between both conditions for other sub-tasks of driving.
The major difference between both conditions of driving without the system (conditions 1 and 4) is in effort supplied for taking into account other vehicles, as traffic density was heavier in the condition in which the effort mentioned in higher.

| Wilcoxon on raw scores | Lateral control | Longitudinal control | Reactivity to dynamic environment | Reactivity to static environment | Itinerary following | Use of controls and equipment | Use of safety indicators |
|---|---|---|---|---|---|---|---|
| Driving without ACC (1) / Driving with ACC | NS | NS | NS | NS | *NC* | **0,03** | **0,01** |
| Driving with ACC / Driving without ACC (4) | NS | NS | NS | NS | *NC* | **0,01** | **0,01** |
| Driving without ACC (1) / Driving without ACC (4) | **0,03** | *NC* | NS | NS | *NC* | *NC* | *NC* |
| NS : no significance; NC : no calculation ; .01 : probability associated to calculated T | | | | | | | |

**Table 7-7 : PSA-TLX -Effort to sub-tasks – group's trend - T Wilcoxon – raw scores**

- *Effort for lateral control*

Most of the drivers judge effort for lateral control as rather low when driving with no assistance. The effort remains rather low when driving with the system, even if it increased (see Fig. 7-16)

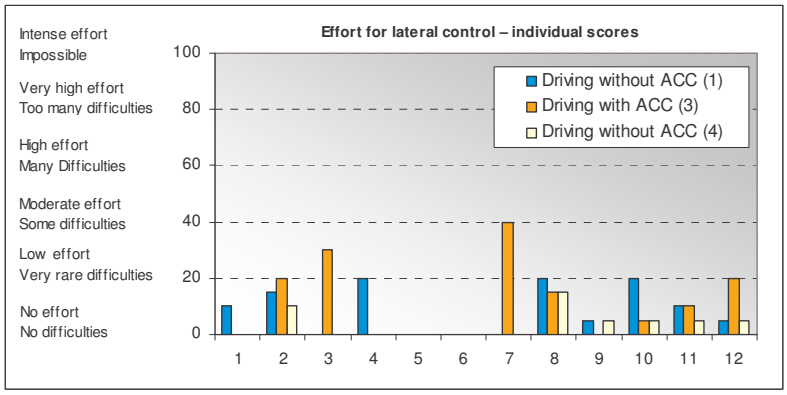



**Fig. 7-16 : PSA-TLX -Effort for lateral control – individual scores**

Nevertheless, one driver mentions low to moderate effort and one driver reports moderate effort when using the system, they felt some difficulties for lateral control.
No driver expresses high effort that can not be maintained.
Around less than a half of the drivers judge effort allocated for lateral control higher when using the system, but the difference is low as the effort estimated is acceptable or optimal.
For the other drivers, no difference is mentioned between driving with the system and driving without assistance: effort is low.

- *Effort for longitudinal control*

Effort for control longitudinal motion is judged rather low or nil when driving without the system (no effort to low effort) see Fig. 7-17.

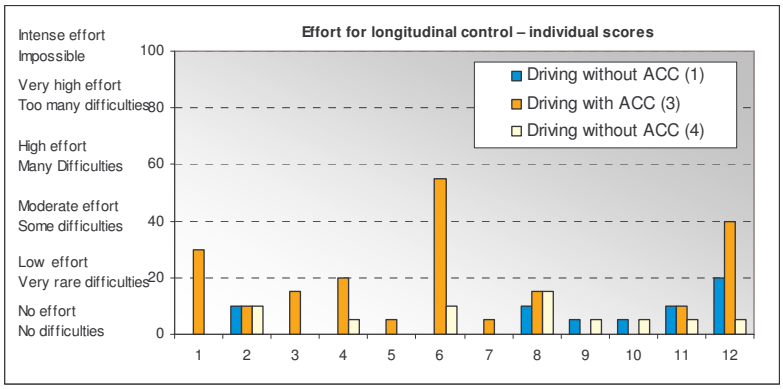



**Fig. 7-17 : PSA-TLX -Effort for longitudinal control – individual scores**

8/12 drivers mention higher effort for longitudinal control when they drive with the system, however the difference is low for most of them and effort is acceptable or optimal.

When driving with the system, most drivers report low effort, except three drivers:

- One driver expresses a low to moderate effort while effort without the system is nil)
- One driver mention moderate effort while low effort is reported without the system
- One driver (number 6) expresses having supplied a high effort hard to maintain, his score reaches the rejection criteria showing that the effort may has a consequence to driver's state or driving performance.

The system's function is directly linked with this sub-task (acceleration/deceleration/braking). Felt effort is linked to confidence in the system and to the need to check and anticipate what the system decides and does ("*need to look at time headway in order not to be slowed down"*; "*problem of detection of other vehicles in curve that are not on the lane"*).

- *Effort for taking into account road users (reactivity to dynamic road environment)*

Scores are globally higher than other sub-tasks, showing that effort required for taking into account road users is higher when driving with the system comparing to driving with no assistance (8/12 drivers), see Fig. 7-18.

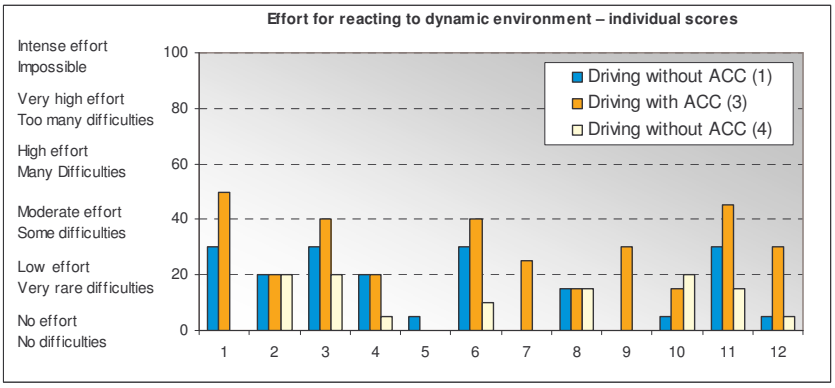



**Fig. 7-18 : PSA-TLX -Effort for reacting to dynamic environment – individual scores**

When driving with the system, variability between drivers appears in effort felt for reacting to other vehicles:

- 2 drivers mention moderate to high effort (without reaching the rejection threshold)
- 5 drivers evaluate the effort low to moderate
- The other 5 drivers express very low or nil effort.

Traffic is dense and the other road users' behaviour has a direct impact on the system's actions. Felt effort is linked partly to potential contradiction between driver's intentions (overtaking the leading vehicle, maintaining an homogenous speed) and the system objective (maintaining distances and target speed).

From this, the driver felt constrained to anticipate system's reactions to get its behaviour compatible with his objectives ("*need to anticipate to avoid being blocked by pulling out*", "high *effort of anticipation to avoid triggering a braking decided by the system in reaction to other user behaviour*", "*anticipation more important for overtaking (respect of the distance with leading vehicle)*").
There is a major difference between both sessions of driving without the system (before and after using the system): effort felt before using the system at the beginning of the experiment is higher than the last condition corresponding to the second time of driving without the system (after using the system). It may be due to higher traffic density and the familiarisation with the vehicle.

- *Effort for taking into account infrastructure and road signs (reactivity to static road environment)*

When driving without the system (before and after having used it), effort to react to static environment is estimated rather low. For almost the half of drivers, effort is higher when driving with the system, but no driver report high effort, only one driver mention moderate effort and some difficulties (see Fig. 7-19)

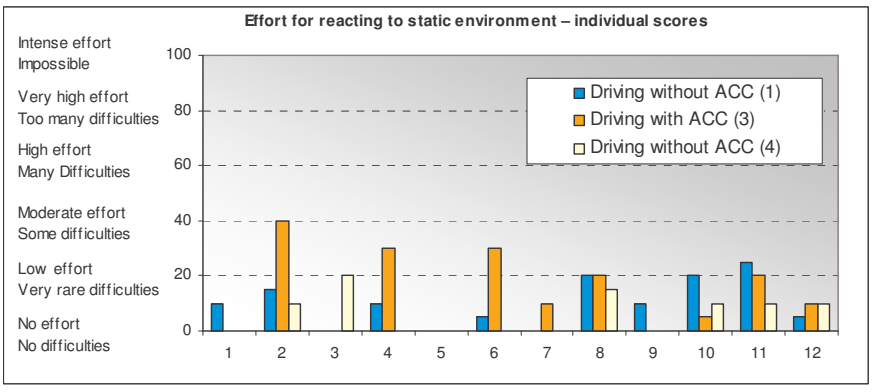



**Fig. 7-19 : PSA-TLX -Effort for reacting to static environment – individual scores**

One third of the group (4/12, drivers number 4, 6, 8, 11) report having neglected this sub-task when driving with the system ("*I did not pay attention to traffic signs*", "*lack of vigilance*").

When driving without the system, most drivers consider effort higher effort when they get used with the vehicle (condition 1) comparing to the second time they drive without the system (condition 4) , but the effort is acceptable or nil, difference is weak.

- *Effort for itinerary following*

Effort linked for itinerary following is judged rather low for the whole trip by most drivers.

Only one driver expresses moderate effort while driving with the system. He says having neglected this sub-task. This is also the case for another driver ("*I did not see the direction sign "Blois"*, "*concentration more on driving than on environment*").
Four drivers estimate having supplied no effort to achieve this sub-task in all test conditions.
Most of drivers (8/12) judge effort for itinerary following as equal in all conditions : few difficulties are met considering that experimenter guides the drivers and the itinerary is familiar to some drivers.

- *Effort for using driving controls*

Effort for using driving control is judged very low when driving without the system. Drivers are used with driving cockpit. They don't feel any particular difficulty. Use of controls is limited due to automatic transmission.
8/12 drivers report higher effort when using the system for using driving controls, where effort is considered low to moderate.
The reason is that drivers knew the conventional cruise control controls. The system's controls are slightly different.
For this condition, two drivers estimate having supplied moderate effort, related to understanding of the system's different modes (activated, inhibited, deactivated, switched off) ("*incoherence of functioning between undergoing and stopped speed/distance management (no pedal vibration)*", "*deactivation of the system is not very clear : hard braking not to be confused with inhibition*").

- *Effort for reacting to in-vehicle state and safety indicators*

Effort for reacting to in-vehicle status and safety indicators is judged very low when driving without the system.
When driving with the system, effort is considered low to moderate by three drivers and one driver mention moderate effort. These have been disturbed by auditory tones emitted by the system and judged not very meaningful ("*beeps don't seem to me very clear: explicit (activation, inhibition, over-passing ..), they reroute attention of the driver*") but also by multiplicity of information displayed on the visual screen (*multiplication of indications with target headway / speed in addition*").

- *Management of effort*

Structure of effort on entire driving task does not change in fundamentals when drivers have the system at their disposal and when they do not have any assistance (no significant difference), the most noticeable difference between conditions without the system (1 and 4) and the condition with the system use (condition 3) is in effort allocated in using driving controls (see Fig. 7-20)

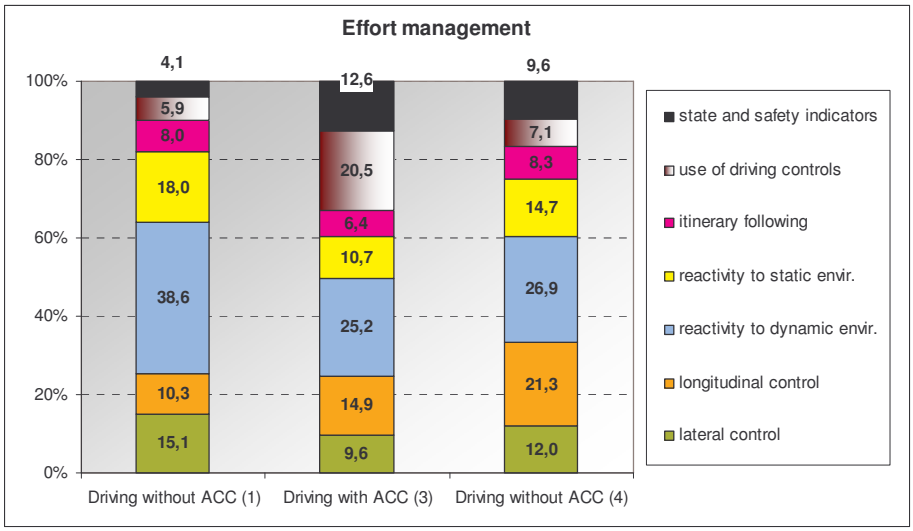



**Fig. 7-20 : PSA-TLX -Effort management**

Differences appear between both sessions of driving without the system in effort for longitudinal control, in reacting to other road users, but differences are not significant.
Some changes are observed reflecting effect other than effect of system use. For example, effort for taking into account other roads users (dynamic environment) is more important in the first condition of driving without the system, comparing to the other two following conditions (with the system and without the system) because of the traffic density more heavy.

- **Efficiency of Compromise between effort and performance**

Global driving

Learning how to drive with ACC generates changes in the way drivers operate the compromise between supplied effort and reached level of performance: driving with the ACC is the only situation for which unfavourable and even very unfavourable compromises can be observed (see Fig. 7-21)

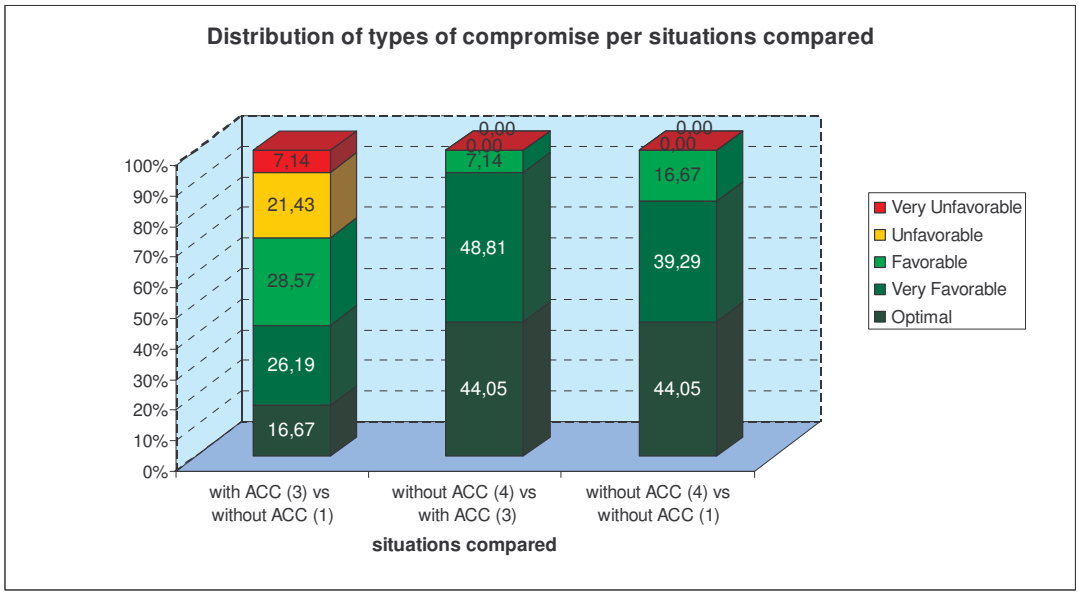



**Fig. 7-21 : PSA-TLX -Changes in compromises efficiency among situations**

Learning to drive with the ACC (condition 3) seems more costly in terms of workload than learning to drive the vehicle (condition 1).
Once this general result obtained, it has to be deepened looking how compromises are distributed among drivers: does this results concerned only few drivers, do drivers always score effort and performance the same (response bias) ?
Furthermore, it has to be also studied how changes are operated on sub-tasks: are all sub-tasks concerned by these changes, are sub-tasks equally costly ?

Individual analysis

For all drivers, driving with the ACC generates a change in operated compromise between supplied effort and reached level of performance (see Fig. 7-22 and Fig. 7-23: and Fig. 7-24: ).
For all drivers, learning to drive with the ACC is more difficult than learning to drive the vehicle (see table 78). The situation may be problematic in terms of workload for one third of the group (very unfavourable compromise for 4 drivers for some tasks).

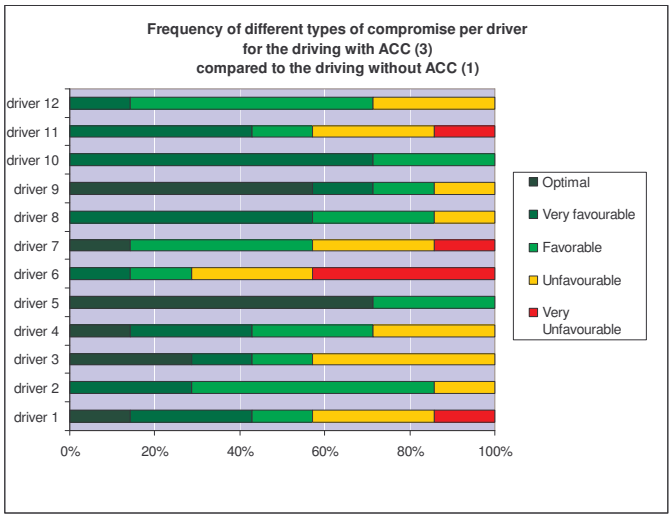



**Fig. 7-22: PSA-TLX -Frequency of different types of compromise Effort/Performance per driver – Driving with ACC (3) vs. driving without ACC (1) - learning vehicle**

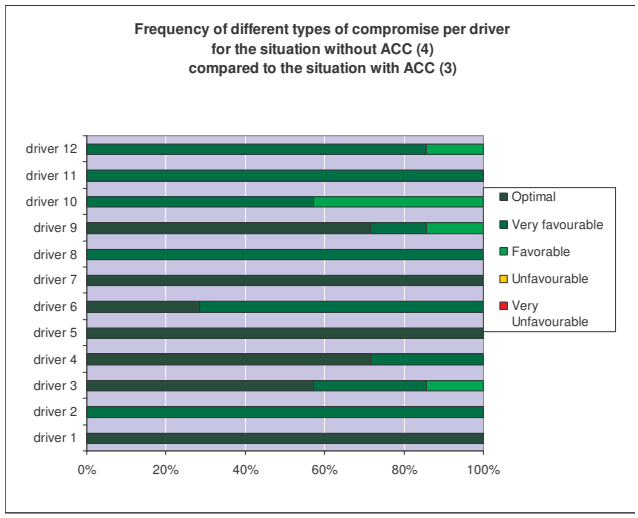



**Fig. 7-23: PSA-TLX -Frequency of different types of compromise Effort/Performance per driver – Driving with ACC (3) vs. driving without ACC (1) - learning vehicle**

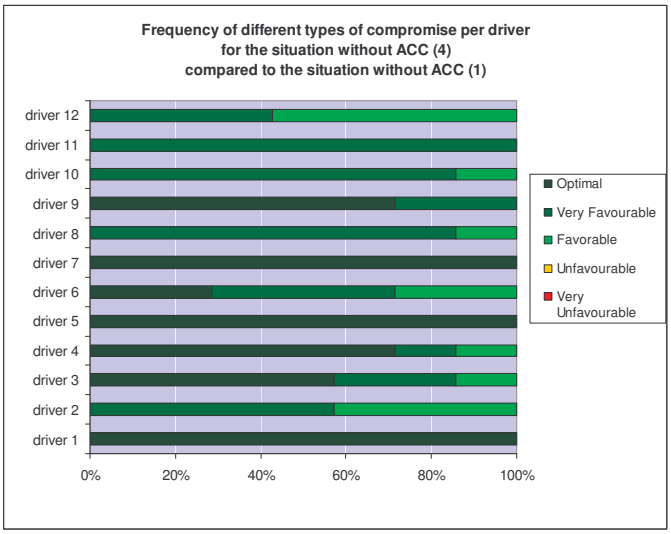



**Fig. 7-24: PSA-TLX -Frequency of different types of compromise Effort/Performance per driver – Driving without ACC (4) vs. driving without ACC (1) - learning vehicle**

A learning effect can also be inferred when comparing the situation when drivers come back to driving without ACC (condition 4) after having used it (i.e. in the graph below "without ACC") to the situation where they drive without it but get used with the vehicle.

Driving without the ACC is estimated less costly (better compromises) when compared to driving with ACC than when compared to driving while learning the vehicle. Learning the ACC is more costly than learning the vehicle for half of the drivers.

Compromise achieved in driving sub-task

Results show that the driving situation does not change the compromise in the same way for each sub-task (see Fig. 7-25 and Fig. 7-26 and Fig. 7-27)

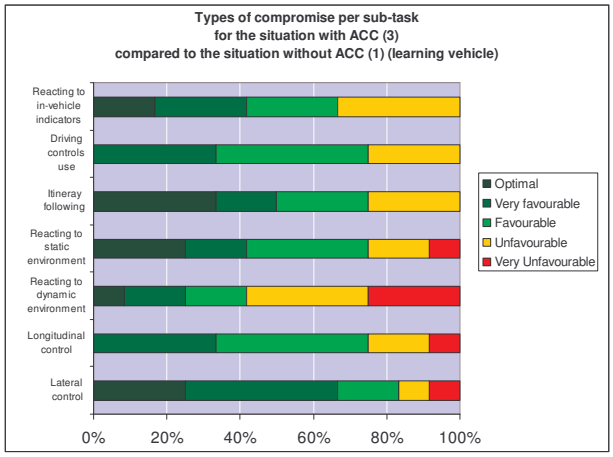



**Fig. 7-25: PSA-TLX -Types of compromise Effort/Performance per sub-task – Driving with ACC vs. driving without ACC (learning vehicle)**

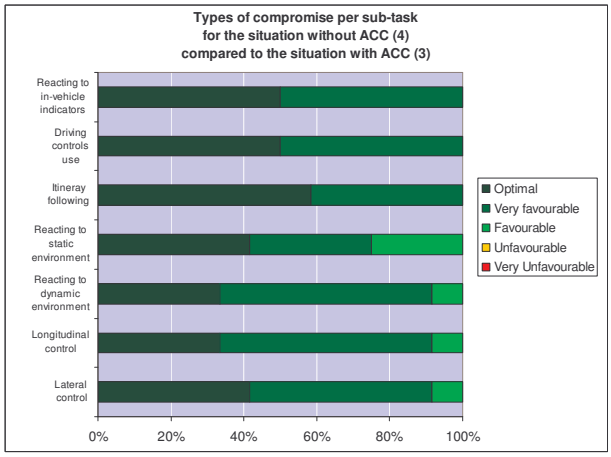



**Fig. 7-26: PSA-TLX -Types of compromise Effort/Performance per sub-task – Driving without ACC vs. driving with ACC**

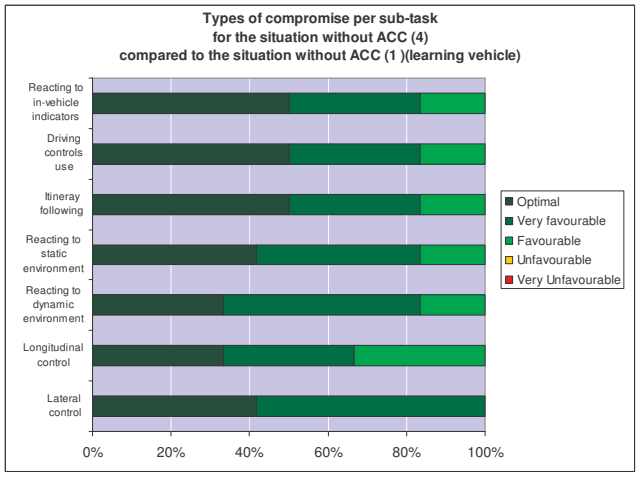



**Fig. 7-27 : PSA-TLX -Types of compromise Effort/Performance per driver – Driving without ACC vs. driving without ACC (learning vehicle)**

Learning to drive with the ACC is particularly costly (worse compromises) for vehicle control tasks and for situational tasks: lateral and longitudinal control, reacting to dynamic and static road situation (taking into account other road users, infrastructure and traffic signs).

When driving with the ACC, reacting to dynamic environment is the task for which drivers operate the worse compromises: they have to learn how to share the task, and at the end of the test-condition, they probably do not manage to find the right compromise.
Other tasks also require either same or more moderate effort for non satisfying level of performance (unfavourable compromises).

This shows that ACC requires a real learning and adaptation behaviour from drivers in order to find a good way to supply a comfortable effort for an acceptable level of performance.

Each sub-tasks is more difficult when learning to drive with the ACC than when learning to drive the vehicle: in this last case, more favourable compromises are observed, with a decrease of optimal and/or very favourable compromises.
Longitudinal control has been reported as a difficult sub-task when getting used with the vehicle: the braking behaviour of the test vehicle surprised some drivers who had to change their usual way of anticipating driving event requiring a braking action.

These results show that the PSA-TLX is sensitive to variations of driving situation demands and allows evaluating learning phase of an ADAS. Drivers manage to report variations between situations and between sub-tasks, showing that skill-based tasks and situational tasks are the most affected when learning how to drive with the ACC which are the tasks concerned by driver-system cooperation issues like decision making sharing, action delegation (operations on pedals), active vs. passive control, supervision of the system. The sub-task consisting in using driving controls is also directly linked with the use of ACC: usual driving controls must be used differently. The sub-task consisting in following itinerary can be affected by interference on attentional management: selective attention focus put on situational tasks and operational tasks.

### 7.2.2 DALI

**Raw scores**

Driving with or without using the system implies medium visual demand and medium effort of attention. The other factors are relatively low (see Fig. 7-28 and Table 7-8).

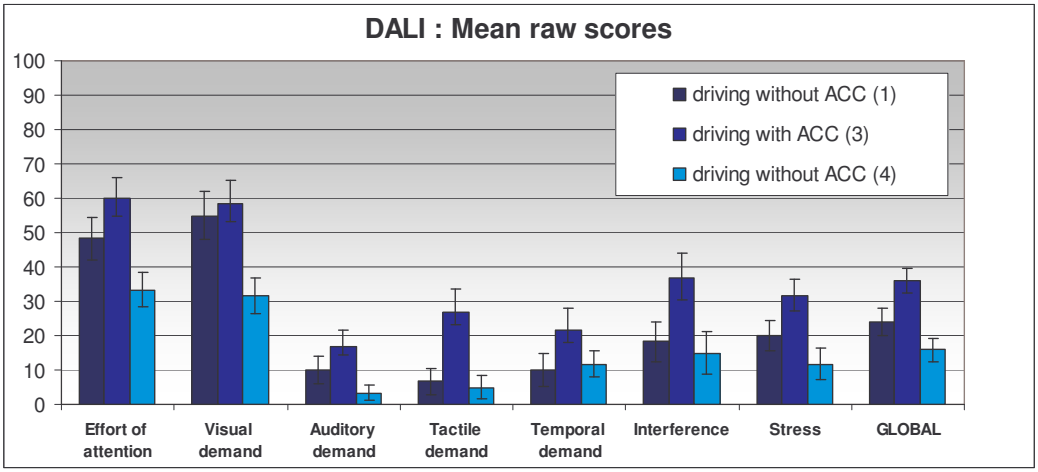



**Fig. 7-28 : DALI - Mean raw scores**

| | Spearman correlation |
|---|---|
| Driving without ACC (1) / driving with ACC (3) | **0,81** |
| Driving with ACC / driving without ACC (4) | **0,93** |
| Driving without ACC (1) / driving without ACC (4) | **0,83** |

**Table 7-8 : DALI - Rank correlation of Spearman – raw scores**

The high and significant rank correlation of Spearman between scores obtained in the different conditions indicates no significant difference between the conditions, when driving with the system (condition 3) and without with the system (conditions 1 and 4).

Wilcoxon test shows significant differences between conditions depending on the factors and the type of comparisons (see Table 7-9):

- After the first session of no-system driving (condition 1), drivers report increased level of interference when driving with the system (condition 3). For the other factors, drivers don't mention any differences: most important factors are effort of attention and visual demand.
- After using the system, drivers feel a decrease of effort of attention, visual demand, interference, stress and auditory demand (for this factor, the level stays quite low in all conditions).
- When both conditions of no-system driving are compared, differences are found on the level of effort of attention and visual demand which significantly decrease.

| Wilcoxon | Effort of Attention | Visual demand | Auditory demand | Tactile demand | Temporal demand | Interference | Stress |
|---|---|---|---|---|---|---|---|
| Driving without ACC (1) / driving with ACC (3) | NS | NS | NS | NS | *NC* | **.01** | NS |
| Driving with ACC / driving without ACC (4) | **.025** | **.005** | **.01** | NS | *NC* | **.005** | **.025** |
| Driving without ACC (1) / driving without ACC (4) | **.025** | **.005** | *NC* | *NC* | *NC* | *NC* | NS |

**Table 7-9 : DALI - Wilcoxon test – raw scores**

## Weighted scores

In this second sub-part, weight of each factor has been computed from the pair wise comparison and is taken into account, so, scores have been weighted.

Effort of attention and visual demand reach a medium level, while the other factors are very low (see Fig. 7-29).

The Spearman statistical index shows very high correlation between scores in all conditions, indicating that no difference between driving with system (condition 2) and driving without system (conditions 1 and 4). The rank correlation is higher than the one computed on the basis of the raw scores (no weighting of factors), see Table 7-10.

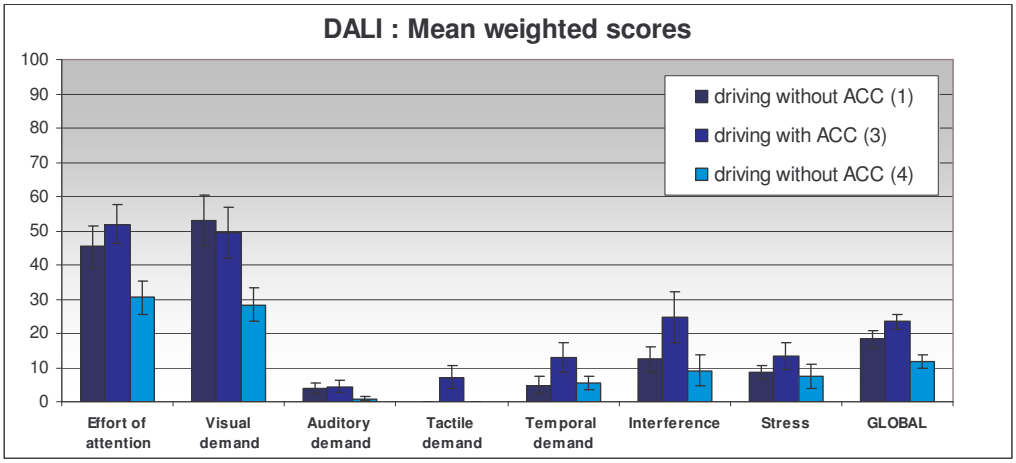



**Fig. 7-29 : DALI - Mean Weighted scores**

| | Spearman |
|---|---|
| Driving without ACC (1) / driving with ACC (3) | **0,92** |
| Driving with ACC / driving without ACC (4) | **0,94** |
| Driving without ACC (1) / driving without ACC (4) | **0,93** |

**Table 7-10 : DALI - Rank correlation of Spearman – weighted scores**

Wilcoxon test shows significant differences between conditions depending on the factors and the type of comparisons (see Table 7-11):

- After the first session of no-system driving (condition 1), drivers report increased level of interference when driving with the system (condition 3). For the other factors, drivers don't mention any differences: most important factors are effort of attention and visual demand.
- After using the system (condition 3), drivers feel a decrease of effort of attention and interference, the level of interference is quite low in both conditions.
- When both conditions of no-system driving are compared (condition 1 and 4), differences are found on the level of effort of attention and visual demand which significantly decrease.

When comparing raw scores and weighted scores, differences are found when comparing driving session with the system (condition 3) and the second session of driving without the system (condition 4). Indeed, the weighting procedure seems to precise the results (less factors show significant differences), as only two factors show differences between both conditions, while raw scores show more significant differences (five factors).

| Wilcoxon | Effort of Attention | Visual demand | Auditory demand | Tactile demand | Temporal demand | Interference | Stress |
|---|---|---|---|---|---|---|---|
| Driving without ACC (1) / driving with ACC (3) | NS | NS | NS | *NC* | *NC* | **.01** | NS |
| Driving with ACC / driving without ACC (4) | **.01** | NS | *NC* | *NC* | NS | **.025** | NS |
| Driving without ACC (1) / driving without ACC (4) | **.005** | **.025** | *NC* | *NC* | *NC* | NS | NS |

**Table 7-11 : DALI - Wilcoxon test – weighted scores**

### Correlation between raw scores and weighted scores

Raw scores and weighted scores are highly and significantly correlated, for all factors (except for tactile demand, as too many nil do not allow correlation computation) and the global score (see Table 7-12).

| Bravais –Pearson Correlation | Effort of attention | Visual demand | Auditory demand | Tactile demand | Temporal demand | Interference | Stress | **GLOBAL** |
|---|---|---|---|---|---|---|---|---|
| Driving without ACC (1) | 0,97 | 0,98 | 0,67 | / | 0,92 | 0,99 | 0,81 | 0,97 |
| Driving with ACC (3) | 0,93 | 0,91 | 0,81 | 0,74 | 0,96 | 0,94 | 0,65 | 0,93 |
| Driving without ACC (4) | 0,98 | 0,95 | 0,94 | / | 0,87 | 0,96 | 0,92 | 0,97 |

**Table 7-12 : DALI - Bravais – Pearson Correlation between raw scores and weighted scores**

### 7.2.3 Evaluation by drivers: final questionnaire

Final questionnaire has been designed to collect the drivers' point of view on the workload questionnaires to fill.
The final questionnaire covers many items: understanding of concepts and explanations, effort required, duration…In total, 10 questions are asked for each administrated tool to compare both tools.
Then general feelings are collected on driver's state and the reasons, and general comments / suggestions.

PSA results of final questionnaire are presented in the detailed technical report (see independent appendix III.1.), and in the section 8.4, so as to have detailed analysis of final questionnaires for three experiments.
In the deliverable, it is more relevant to have the analysis of all data collected in the three experiments, as the aim is to compare the three tools.

Main PSA results of final questionnaire are presented here:

No difference is observed between both PSA-TLX and DALI in:

- Duration of filling in : "quick"
- Formalism :" well"
- Questions understanding :" well"
- Explanation given : "sufficient"
- Evolution of effort (learning effect) : "easier" or "equal complexity"

Notable differences between DALI and PSA-TLX:

- DALI: Easier for scoring (75%vs 68%)
- PSA-TLX: concepts easier to understand (100% vs. 75%)
- DALI: less effort ( 50% vs. 75% for "moderate level") for most drivers except for one driver who report high effort while no high level of effort is mentioned for PSA-TLX.

Small differences between DALI and PSA-TLX:

- DALI: Easier for scoring (75%vs 68%)
- PSA-TLX: a bit more "exhaustive" (75% vs. 60%)
- PSA-TLX: written explanations easier to understand (92% vs. 75%)
- PSA-TLX: explanations given by experimenter judged less "necessary" (68% vs. 82%)

<u>General comments of drivers :</u>

DALI:

- Need of examples to define and understand concepts to evaluate
- Difficulty to distinguish "temporal demand" and "interference"
- Difficulty to make evaluation without any reference situation for comparison

PSA-TLX:

- Scale is too large (too many choices)
- Scale graduation is too "subtle" (preference for 10 by 10 points)
- Questionnaire too "heavy" in terms written explanations and questions

## *7.3 Discussion on PSA results*

A road experiment was conducted with an ACC to evaluate two workload subjective tools – the PSA-TLX dedicated to evaluate driving workload (driving performance and effort of driving, driver's state) and the DALI dedicated to evaluate mental and perceptual demand.

Results showed that both tools are sensitive and complementary, as they are dedicated to different purpose and they evaluate different workload components of different natures.
Drivers reported low or moderate effort to fill the questionnaire (for one driver, DALI requires high effort, while no one reports high effort for PSA-TLX).
Concept and explanations did not present any particular difficulty, and scoring is relatively easy considering the scale.

Nevertheless, the administration of both subjective methods may induce interference:

- both tools present major differences in evaluated activity, scoring process: these differences may induce **errors** for the driver
- both tools present some similarity in evaluated concepts (e.g. effort of attention, interference…), this likeness may induce **confusion**

The use of only one method seems to be more appropriate, as their concurrent administration could lead to "parasite" difficult to control by the experimenter.

Results showed that DALI would be more relevant for HMI evaluation, if the choice between several HMI is required as perceptual and mental demands are evaluated, while PSA-TLX would be more appropriate to decide if a system is rejected or not, considering that the impact on driving activity in terms of driving performance and effort is estimated.

# 8 Comparison between different Methods/metrics

After having described experimental settings and protocols that were followed during experimental sessions, a comparison between tools used was carried out on the basis of the results collected with Final Questionnaire (see independent appendix IV.2). Subjects were requested to fill in this questionnaire after having administered the specific subjective workload assessment tools.

## *8.1 Description of procedure*

On the basis of results obtained in Final Questionnaire analysis, some graphs representing the overall results have been built: these graphs can easily represent every single question with the corresponding choices for every questionnaire. When filling in Final Questionnaire, subjects were well informed about the aim of that further, time requesting, questionnaire: they were told that the real aim of all experimental session was to evaluate the potential of the three questionnaires for workload detection and reporting, in or after, a driving simulation.

When making comparisons of partners' results, a first important point not to be forgotten is that experimental samples are quite little and that not the same number of people administered the same questionnaires, and not referring to the same workload situation experienced. There was an overall number of 47 people involved in the whole experimental design, divided in 3 locations and in a bigger number of experimental conditions. Moreover, and this is the most remarkable point to be taken into consideration when comparing results, number of responding subjects are not the same for every questionnaire considered.

PSA utilized only two questionnaire instead of three (PSA-TLX and DALI), while INRETS administered all the three questionnaire but using two different experimental conditions (in the first case DALI and PSA-TLX were used, while in the second DALI and BMDMW were assessed). The global scores (choices) regarding final questionnaire have therefore been built, and these scores have been transformed in percentage, to allow exploitable comparisons: we have always to keep in mind that there were subjects evaluating sometimes 2 and sometimes 3 questionnaires (a difference that influence of course also decision making and preference assignment processes), and that a percentage could be a result merging from a restricted group of subjects.

## *8.2 Criteria for comparison*

In workload assessment literature some recurrent criteria for comparisons can be easily encountered. These criteria have been frequently used to compare some different tools on the basis of some merging characteristics. Among them we find ease-of-use, sensitivity, diagnosticity, transferability.
All this criteria, when utilized, need a clear definition and a method to be assessed (difficult to establish how much a tool is "diagnostic" if it is difficult to assess workload factors).

In this overall experimental phase, drivers were requested to concentrate their efforts after driving on workload experienced and to try to estimate it and to rate it on a particular assessment scale. After having done this operation for three (or two times) the real objective of the design came out: to compare the instruments just administered. Subjects were not asked if tools were diagnostic or sensitive, or if tools were adapt to clearly assess "mental workload" (a concept that can of course encounter many different and personal acceptations). Subjects

were asked if tools were well-presented, if questions were clear and well-made, if tools were time expending or if the underlying concepts were understood, together with for example if questionnaire would have expressed all their personal feelings.
This kind of evaluations is of course more useful than an "objective classification" that mask personal perceptions and comments because of its "scientificity" difficult to be found by a subject.

Final questionnaires items (questions) are therefore the real criteria of assessment, useful to understand what people thought of a just experienced tool administration. Almost every single question of Final Questionnaire address a different important concept to have a preliminary sketch of potentialities that the single tool could provide when they will be utilized to assess workload after a driving sequence interacting with an adaptive and integrated interface.

## *8.3 Assessment of the considered (single) methods*

Here below we report a brief discussion for every single question of Final Questionnaire, by observing answers collected by partners and showing in a graphical manner the overall results.

Question 1:

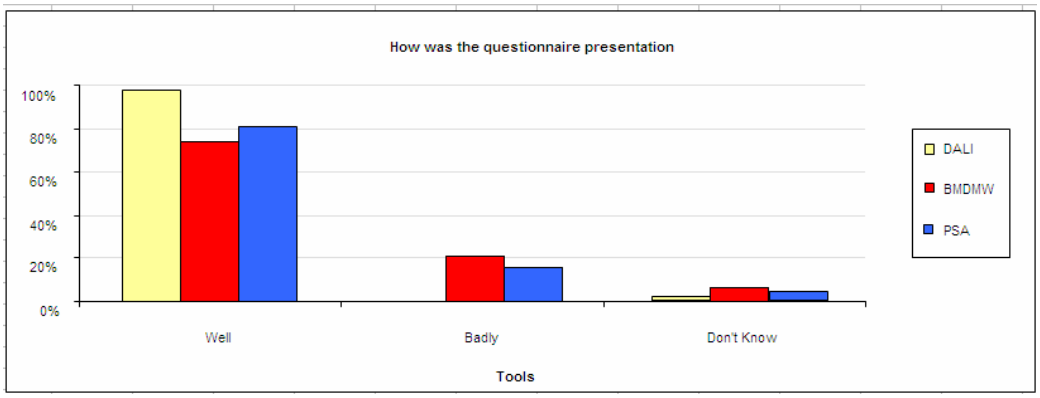



**Fig. 8-1 : Questionnaire presentation – all experiments**

In this question DALI received the best answers because nobody retained it as "badly" presented. Although for the other tools some kind of this responses appear we can conclude that all the three tools were well presented. We remember that these percentages come out from different samples of respondents, therefore 20 % referred to DALI means more people than the same percentage referred to PSA-TLX or moreover to BMDMW.

Question 2:

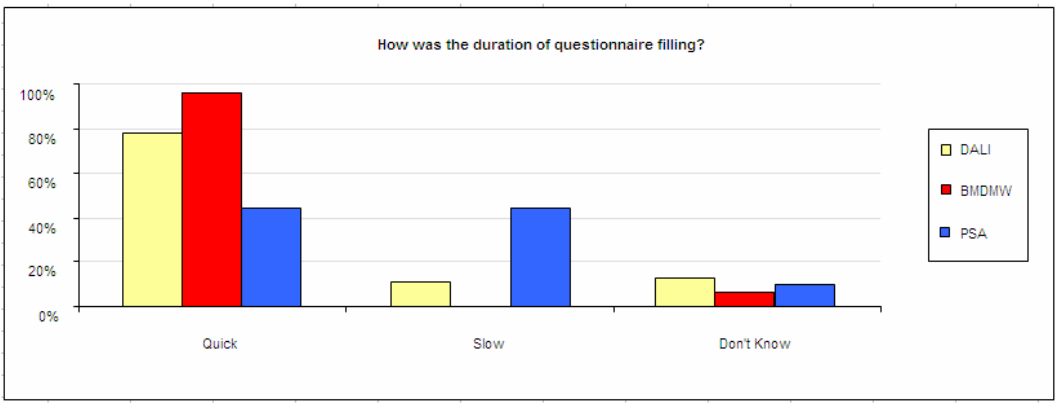



**Fig. 8-2 : Duration of questionnaire filling in – all experiments**

DALI and BMDMW are reported as tools quick to be filled in with a low percentage (10%) of "Slow" responses reported only for DALI. This percentage results increased for PSA-TLX, which is reported as of course a tool very long to be administered (almost half of participants).

Question 3:

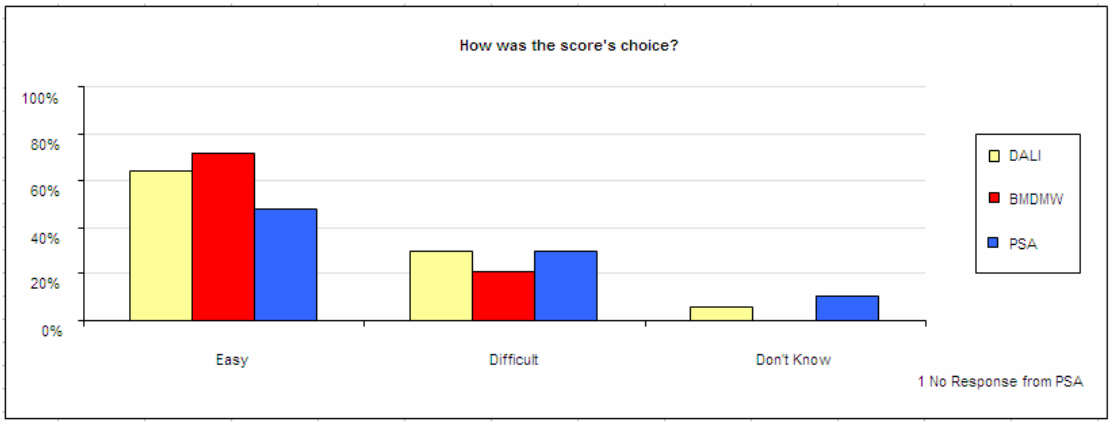



**Fig. 8-3 : Score's choice – all experiments**

In this question BMDMW delivered the easiest score's choice (95% of "Easy" answers in CRF experiments but 40% in INRETS ones). Also in DALI score's choice resulted very easy (100% in CRF answers and near 70% for other partners.
PSA-TLX results sounded ambiguous, as subjects from PSA and INRETS judged it easy (75% and 60% respectively) while for CRF subjects it resulted difficult for nearly 75% of participants (they thought that PSA-TLX scale is difficult to use because too wide). PSA-TLX asks also to answer thinking at a reference situation that could be difficult to imagine when in simulator trials.

Question 4:

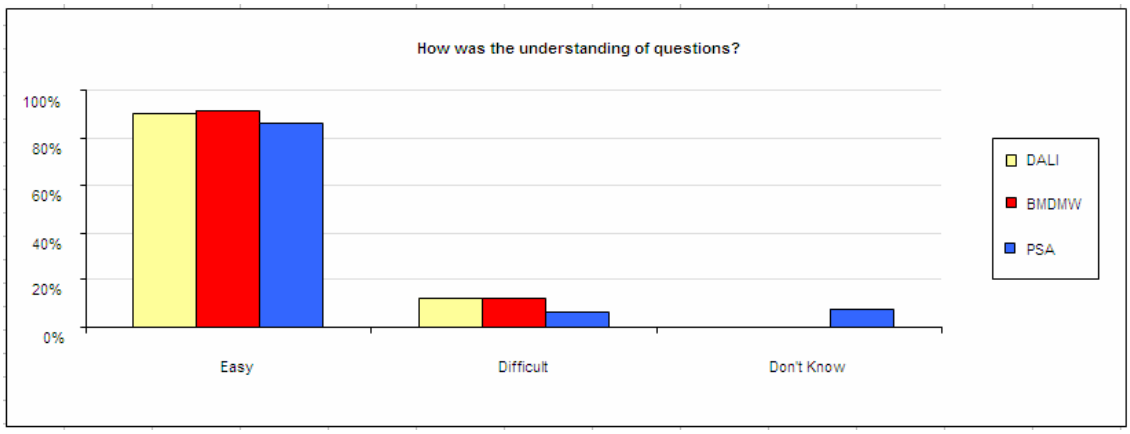



**Fig. 8-4 : Questions understanding – all experiments**

PSA-TLX resulted easy not for all participants: CRF found that some drivers (15%) it was difficult while for INRETS all the drivers found it easy (100%). DALI obtained the best results (between 85% and 100%) of easiness. BMDMW was found difficult by 20% of drivers in INRETS experiments, while for CRF subjects (100%) it was retained easy.
We can affirm that questions are understood by almost all the drivers, although to understand them it could be requested more time in some cases (see also results from question 3 above).

Question 5:

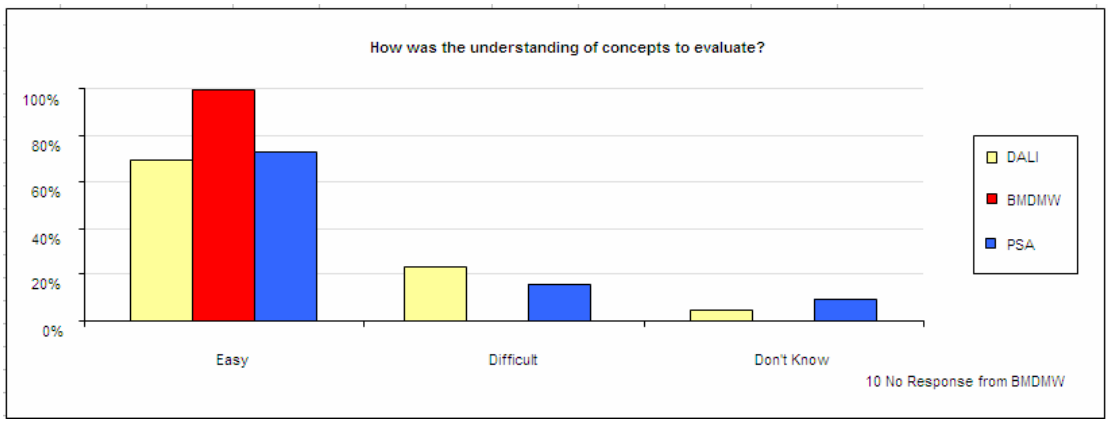



**Fig. 8-5 : Concept understanding – all experiments**

While BMDMW (data only from CRF) has good understandability of concepts evaluated, some difficulties are reported for PSA-TLX (in INRETS trials, 30% of drivers marked this alternative as well as for a third of CRF subjects). Also DALI presented some "Difficult" choices especially in INRETS trials (35%) and in PSA ones (17%). BMDMW results the best tool from this point of view but available data are not so many as well as for other tools. In INRETS trials there are 10 no response, meaning that during one experimental session (more precisely the one with DALI vs. BMDMW) drivers did not express opinions about this tool. For the other two tools the results are quite the same.

Question 6:

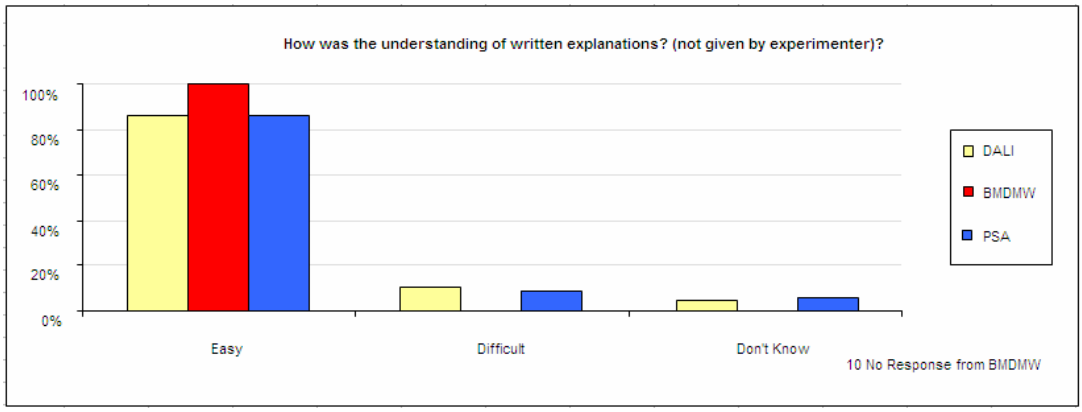



**Fig. 8-6 : understanding of written explanations – all experiments**

PSA-TLX furnished easy written explanations for 90% of subject in PSA experiments. This percentage is the same in INRETS results, while it is lower in CRF trials (70% with 20% of "Difficult" choices). DALI obtained better percentages in INRETS and CRF trials (85% and 93% respectively), while BMDMW obtained 100% of "Easy" choices but only in CRF trials, as from INRETS again there were 10 no response (=100%). All the percentages after all demonstrate a good format of written explanations, understandable and easy.
Question 7a:

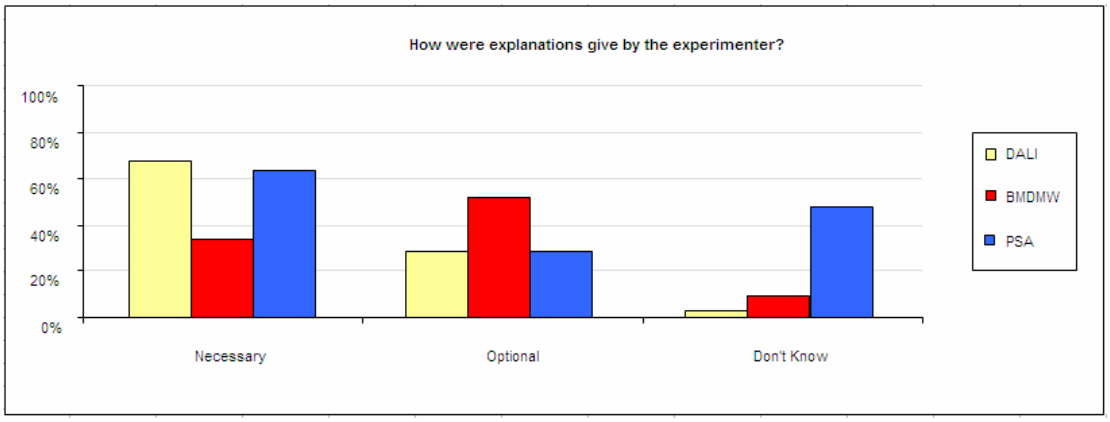



**Fig. 8-7 : Explanations of experimenter (necessary)– all experiments**

This question delivered very different results from the different experimental situations. While for example in INRETS all the questionnaires were retained difficult without experimenter suggestions (that were reported as "necessary" from about 70% of subjects for every tool), in CRF this percentage fall down to 45% percent for DALI, 60% for PSA-TLX and 10% for BMDMW. In PSA trials, "necessary" choices were quite the same for both the tools, although it was found a significant difference during the analysis.

These results seem to be impossible to be interpreted in a univocal manner: probably the most important reason of this heterogeneousness is the content of instruction itself. Instructions were not "coded", but were subjectively provided. The different effects obtained on drivers perception is function of what has been said and which added information have been provided, that means that different outcomes could have been depended on individual attitudes and choices made by experimenters.

Question 7b:

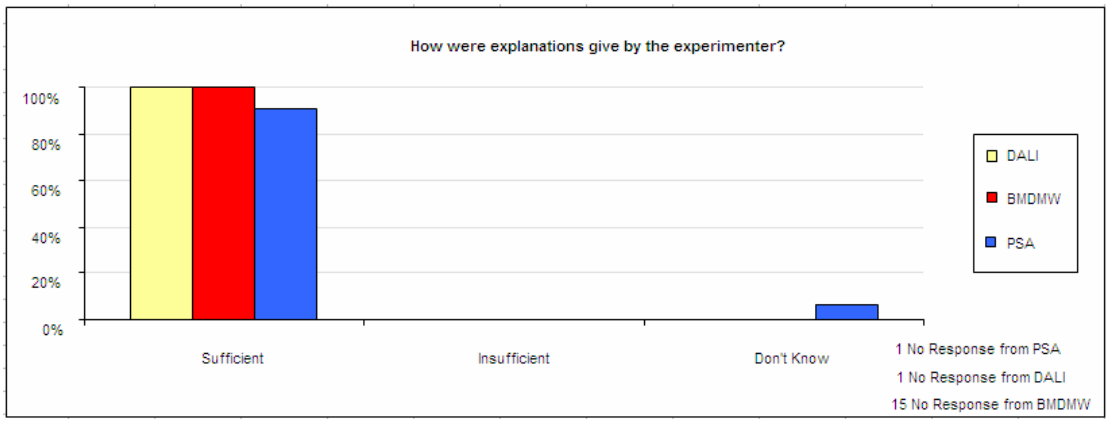



**Fig. 8-8 : Explanations of experimenter (sufficient) – all experiments**

The different outcomes of previous question are not present here: all the explanations given were considered as sufficient, without need of improvements. The "insufficient" alternatives have not been chosen, proving the fact that written explanations together with verbal precisations can assure a great understand between subjects.

Question 8:

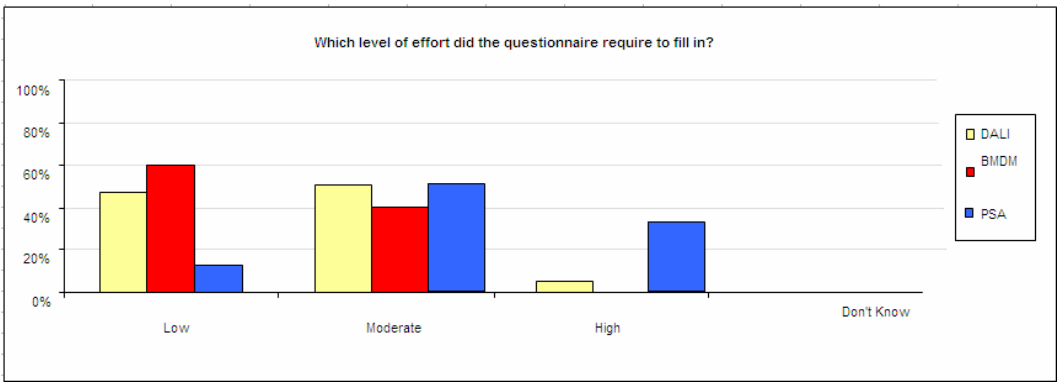



**Fig. 8-9 : Required effort to fill in questionnaires – all experiments**

This question have provided as above for question 7, different results: PSA-TLX for examples have been retained high-effort requiring for 40% of INRETS drivers and 60% of CRF drivers, while from PSA data only DALI have been perceived in this way by a little percentage of drivers (8%). BMDMW and mainly also DALI divide preferences among "Low" and " Moderate" labels.

Question 9:

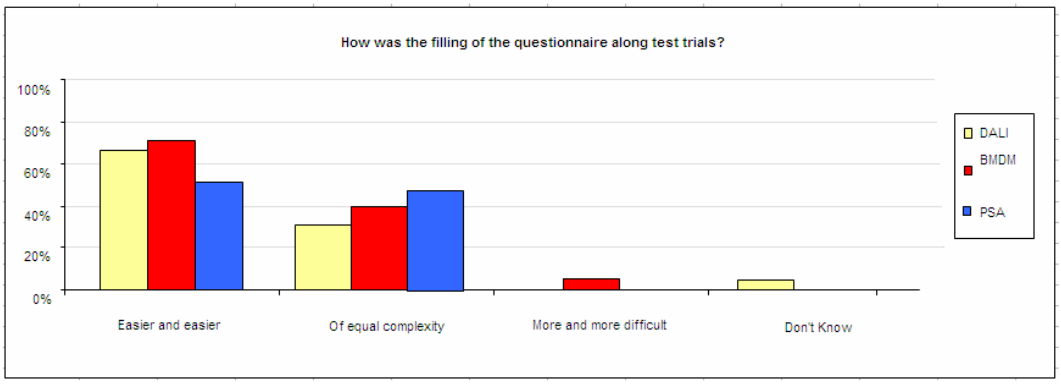



**Fig. 8-10 : Evolution of effort required to fill in questionnaire – all experiments**

In general, fill in the questionnaire becomes easier and easier, although with different percentages that satisfy this condition. The most stable tool is PSA-TLX, which is reported with the same complexity in almost half of the cases.

Of course experience in compiling the questionnaire becomes greater from the second time yet: familiarization is of course a facilitating construct when dealing with subjective assessment comprehension. Nobody reported tools becoming the more and more difficult except one person in INRETS trials.

Question 10:

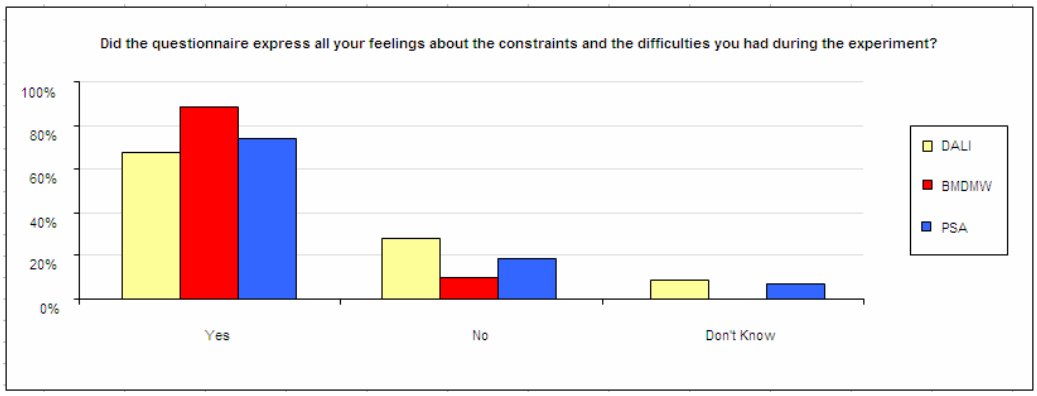



**Fig. 8-11 : Exhaustivity of questionnaire – all experiments**

All the questionnaires seem to express, more or less at the same level, drivers feelings about constraints and difficulties encountered. These results are achieved with different partial outcomes from partners that counterbalanced each other when the overall scores are computed.
DALI is reported as a satisfactory tool with almost the same percentage in the three cases. The same for negative choices, although these ones are a bit bigger in PSA trials than in other ones.
Regarding BMDMW, in CRF experiment it was the most exhaustive (90% of "yes" responses) while in INRETS results this percentage stop at 60%, being the tool with the highest percentage (30%) of "no"answers".
PSA-TLX reported very successful and excellent results in some tests, as in PSA-TLX trials or in INRETS ones. On the contrary, in CRF answers it revealed itself as the less expressive,

with a slight difference respect to others. This is the reason because it does not receive the best result in the global graph.

**General section:**

Here below there are reported results obtained from the last questions of Final Questionnaire, about subject feeling and individual state after the experimental session.

Question 11:

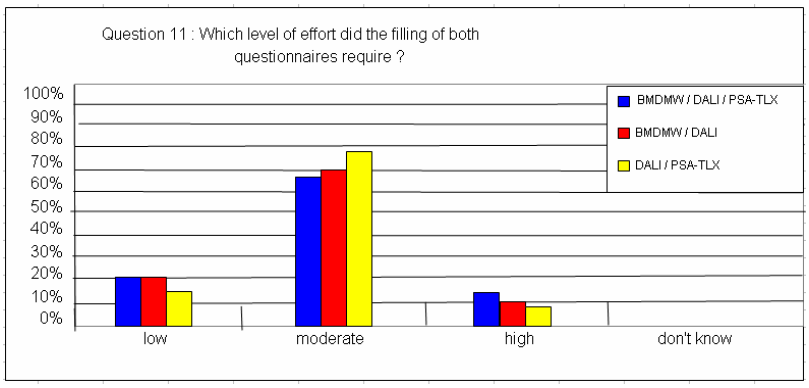



**Fig. 8-12 : Effort required to fill-in questionnaires – all experiments**

More than 70% of subjects reported a moderate effort after questionnaires filling, while for only a smaller group this effort was high or low. Of course to fill in three questionnaire (two for INRETS and PSA subjects) is a long time consuming task, which becomes more effort-requiring if one of more questionnaire are not so clear and easy to be compiled (as referred for other tools). Above, a general graph has been built and presented, but because of differences in experimental conditions, three groups have to be distinguished, on the basis of number and kind of administered questionnaire. The first group (Group 1, in blue) represents CRF subjects, the only ones that administered 3 questionnaires; the second one (Group 2, in red) represents half of INRETS subjects, that filled in DALI and BMDMW questionnaires; the last group (Group 3, in yellow) deals with subjects that administered DALI and PSA-TLX questionnaire therefore with PSA subjects and the remaining part of INRETS subjects.
Looking at Group 2 and Group 3 data, it can be affirmed that the situation can be considered quite the same as for Group 3, as there are not big differences. These results seem to suggest that effort required does not depend from the number of tools under evaluation

Question 12:

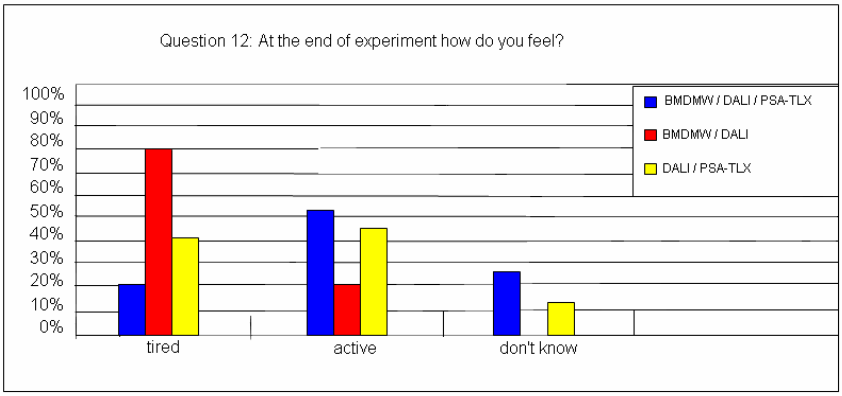



**Fig. 8-13 : level of fatigue at the end of experiment – All experiments**

From the graph above it can be affirmed that almost 50% of people referred to be tired at the end of experimental sessions. Nevertheless, this global result could not account of great differences for single groups result. Group 2 for example, presents a very interesting result: 80% of people referred to be tired after the experiment. Considering that there were only two tools evaluated in this condition, independence of number of tools from tiredness generated seems here confirmed. In fact Group 1 shows people more active although 3 tools were administered.

From Group 3 also, it can be noted that tiredness does not seem to depend on number of tools evaluated but moreover on the type of tool under evaluation: subjects administering PSA-TLX and/or DALI referred to be more tired than when there were 3 tools under assessment. More precisely DALI seems to be more connected with tiredness generation: there were in fact more people tired when DALI was assessed with BMDMW than with PSA-TLX for example.

Question 13:

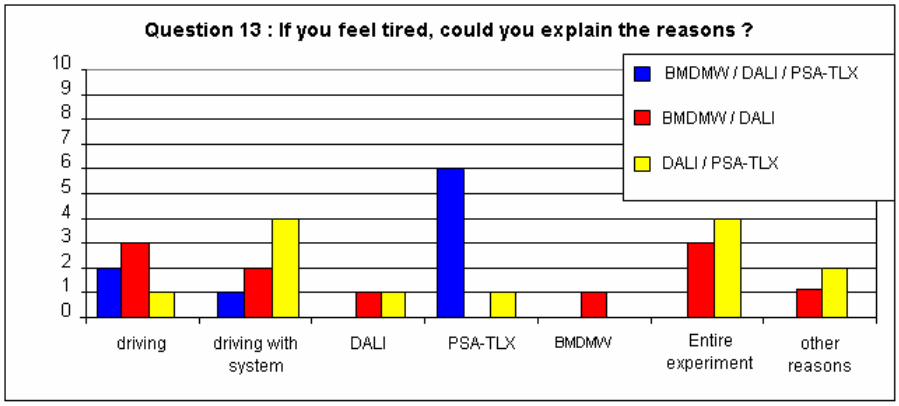



**Fig. 8-14 : Reasons of tiredness – all experiments**

In this question multiple choices were allowed when answering and result are report with absolute values instead of percentages. People indicated possible reasons of their referred tiredness. From these answers it can be seen how a driving situation is by itself stressful and effort requiring: driving and driving with systems account for almost a half of reported answers. It is interesting to underline the fact that of the three tools being evaluated, only PSA-TLX have been reported as a cause if tiredness by more than 20% of people. DALI and BMDMW have been also chosen among other possibilities but by less than 5% of people.

The first graph above is in relation with CRF administration of all the three tools: as it can be seen, except from the driving situation and the use of systems, PSA-TLX (Questionnaire B) have been indicated as one of the reason for being tired (of course not the only one, because multiple choices were allowed). PSA-TLX have been indicated as cause of tiredness mainly by people who administered all the three questionnaire, while people that dealt with it in the last condition (DALI-PSA-TLX) did not choose it as cause of tiredness.

## *8.4 Experimental evaluation of some selected methods*

In this section some criteria are proposed to look at the collected data. As previously said before presenting final data, this criteria were merged from the type of comparisons that experimenters wanted subjects to perform, i.e. from question of Final Questionnaire, that indicate what subjects had to think of when assessing and mentally comparing the three (or two) tools of interest.
By proposing these criteria, the idea is to consider tools utilised in this experimental phase from the point of view of their "natural" end-users (drivers) more than from a traditional scientific point of view, for which some results have been already reported in others deliverables (as in D 2.2.1 "Review of existing techniques and metrics for IVIS and ADAS assessment").

Here below, criteria chosen to have a sort of response about applicability of the tools utilised in future, more complex, experimental sessions, when an integrated interface will be the system under evaluation not only from subjective mental workload perspective.

1. Duration

Subjects liked DALI and BMDMW as quick assessment tool, while for half of the subjects that administered PSA-TLX, assessment procedure was retained slow. Considering that DALI was administered by the biggest number of subjects, this tool shows the best result.

2. Applicability (score choice)

Score choice was reported as easy from two thirds of subjects more or less for every tool examined. They are therefore equivalent under this point of view, although for PSA-TLX again this result is weaker.

3. Clearness (understanding of questions and concepts evaluated)

All the questionnaires have shown themselves to be clear in questions presentation and in concepts that subjects are asked to assess. This result is very strong for all the three tools and there are no remarkable differences. Nevertheless, it has to be noted that for example there is a great amount of missing answers for BMDMW evaluation, and that the actual result could have been stronger or weaker.

4. Effort required

From the point of view of effort required to fill in questionnaires, a sort of order can be traced: BMDMW obtained the best result (60% of "low" rates), because nobody chose "high" alternative for this tool. DALI obtained a good intermediate result, with people dividing more or less equally their results between "low" and "intermediate" alternatives. More effort have been reported to fill in PSA-TLX, with only 15% of people reporting "low" effort, 50% "moderate" and 35% "high".

5. Completeness

In this proposed criterion we include answers given to the last question of first section of Final Questionnaire, related to personal feelings. The best result in this question was obtained by BMDMW, with only 10% of people affirming that the questionnaire did not express all their feelings about difficulties experienced. The other two tools were retained exhaustive, with a slight difference respect to the first (about 75% for PSA-TLX and 65% for DALI). All the questionnaires have covered possible feelings coming out from a driving experience. We could affirm that subjects have confirmed construct validity of the three questionnaires.

In Deliverable 2.2.1 "Review of existing techniques and metrics for IVIS and ADAS assessment" some subjective methods were cited and reviewed. DALI and PSA-TLX were described and a short list of advantages and drawbacks were compiled for every method discussed.

Concerning PSA-TLX, some objectives were listed, representing some characteristics of interest for the tool. After Task 2.2.6 experimental phase, some of them have been surely reached, while some others have only partially been respected, as drivers' answers to Final Questionnaire have confirmed.

In D 2.2.1 the purpose to have a tool with the following characteristics were considered:

- Easy to use for experimenter and for drivers
- Cost-effective
- Not intrusive
- Able to identify components of workload
- Able to identify impact on driving task
- System-independent

Drivers answers to Final Questionnaire (and also data collected with this tool about workload experienced) have confirmed PSA-TLX as a system-independent tool, diagnostic of workload components and able to make quantifiable impact on driving task of some interactive processes (with an ADA or a IVI system). These qualities have been confirmed but together with a shared opinion among drivers that all the assessment process is too time-requiring and not always easy to understand and therefore to use. The first two objectives of the list above have been therefore only in part confirmed.

For DALI questionnaire, for which some issues for further research were listed, it can be said that from these experiments again there was a part of subjects that still showed some difficulties in comprehension and in differentiation of Factors, and maybe difficulty to identify target of assessment (system in use or driving performance?) was still present.

Of course this experimental phase represented a good and rare opportunity to collect "the same data" with different assessment tool. By administering the same tool referring to the same driving situation, subjects could then be able to compare and evaluate easiness and friendliness of different subjects.

BMDMW has shown great strength points that could be improved in some future assessment phases that will be of course necessary to have further validating data. Also in trials only a few part of participants dealt with this questionnaires. Percentages referring to BMDMW have been calculated on a smaller number of respondents, therefore further data will be needed.

PSA-TLX has in general demonstrated and confirmed its strength points, although also some drawbacks have been confirmed. This questionnais have been described as too long and sometimes of difficult understanding, maybe because some information and parameters are automatically processed by drivers and therefore difficult to be evaluated "off-line", as control of the lateral motion.

DALI has confirmed its good result of practical and easy to use assessment tool, although there are sometimes difficulties in understanding concepts being evaluated, and although it is not clear if weighting phase effectively help to identify causes of workload.

# 9 Conclusions: Specification of suggested method/metric to be used in the AIDE test regime

The aim of task 2.2.6 work was to evaluate the potential of three existing methods – BMDMW, DALI and PSA-TLX- developed by different research department to evaluate workload for car-driving context. Among these workload subjective methods, two have already been tested on road with different case studies.

To reach this goal, three experiments have been conducted in different test environments (virtual reality simulator and road), with three different case studies covering ADAS and IVIS (FCWS, ACC, route guidance system), to fit with AIDE project context.
Work conducted inside task 2.2.3 was to attempt to give elements to choose only one tool as being the "best", and to define "workload" during driving with using IVIS or with ADAS.

The evaluation of the three questionnaires was based on driver's point of view to assess the "usability" of the questionnaire (e.g. effort required) and estimation of methods' sensitivity, in order to assess the interest of each questionnaire in IVIS/ADAS design.

Experiments have been built taking into account methodological and practical issues: constraints for the driver, time constraints for the experiment, potential interference between tools which could affect each tool evaluation.

Results showed that the three questionnaires:
- evaluate different workload components of different natures (perceptual demand, mental demand, driving workload, driver's state),
- are dedicated to different purposes : HMI design (DALI) or subjective evaluation of risk in driving task (PSA-TLX and BMDMW)
- Present different level of sensitivity depending on test environment and task's demands: DALI and BMDMW seem to be sensitive in simulator and road environment, while PSA-TLX is sensitive on road.
- Different practical constraints : PSA-TLX is longer to administrate than DALI or BMDMW
- Different usability: in INRETS and CRF experiment, some drivers report serious difficulties to fill and understand PSA-TLX and in PSA experiments, some drivers report difficulties to fill and understand DALI. This result means that experimenter training and skill is essential to administrate a questionnaire properly. From the way the questionnaire is administrated depends the scoring and the interpretation.

Consequently, only one "best" method has not been recommended as the three methods are complementary, but the area of use for each method has been specified to help the choice between them depending on the questions.

Recommendations have been provided for each method by specifying the purpose, the requirements, the advantages, the drawbacks and the limits of the tools.
They are presented in Table 9-1, Table 9-2 and Table 9-3.

| **Behavioural Markers of Driver Mental Workload (BMDMW)** | |
|---|---|
| **Purpose** | To measure Mental Workload through a subjective measure based on ratings of behavioural patterns related to driving tasks. |
| **General description** | This tool is a questionnaire using a 5-point Likert scale ("never" to "very often"). It consists of 27 general items plus 7 specific items related to Urban Environments |
| **Advantages** | *Sensitive to differences* among conditions but not as sensitive as DALI |
| **Advantages** | • Sensitive on road and in VR Simulator to differences among conditions (task's demands)<br>• *Quick administration and data processing and analysis* (automatic statistical analysis Global score + factors of workload)<br>• *Flexibility*: tool adaptable to different scenarios (by adding or removing specific items) |
| **Drawbacks** | • *Understandable concepts* and *efficient explanations* |
| | • **Requires more experimental feedback and validation in order to refine the single factors** |
| **Drawbacks** | • Mixing qualitative and quantitative values for the scale<br>• Global score not enough informative |
| | • First application in the present research; further validation is needed |
| **Limits** | • Same answer can be explained by different natures of load |
| **Minimum number of participants** | • Not less than 12 participants |
| **Modality of administration** | Self-administered questionnaire. It is important that the tool administration happens immediately after the conclusion of session under evaluation |
| **Instructions** | Written instructions reported in the questionnaire seem to be sufficient |
| **Mean time to fill the questionnaire in** | First administration around 5 minutes, learning effect (third administration) around 3 minutes |
| **Type of analysis** | • For less than 30 participants Wilcoxon test is recommended |
| **Type of analysis** | • For more than 30 participants Student test is recommended |
| | • |

**Table 9-1 – Recommendations for BMDMW questionnaire**

| **DALI** | |
|---|---|
| Purpose | To evaluate driver's mental workload through a subjective measure based on rating factors dealing with perception, cognition, temporal demands and state of the driver related to the tasks. |
| General description | This tool is a questionnaire displaying 7 factors (possibility to add or to withdraw factors) using a 6-point scale from 0 to 5 : Effort of attention, Visual demand, Auditory demand, Tactile demand, Temporal demand, Interference, Situation stress.<br>In a second optional step, it is possible to conduct pair-wise comparison factor by factor in order to weight each factor value according to the results of these pair wise comparisons. |
| Advantages | Sensitive to driver's workload differences among conditions<br>Sensitive on road and in driving simulator to differences task's demands (driving with IVIS or ADAS)<br>Concepts and process easily understandable for the driver<br>Quick administration and quick data processing for the experimenter (software available to enter directly data by the driver and to compute the weighted factors values)<br>Quick and easy analysis (low number of factors)<br>High correlation between raw scores and weighted scores would allow to make this tool lighter by suppressing pair wise comparison for a quicker administration.<br>Possibility to have a global score in addition to specific scores related to visual, auditory,…workload.<br>This tool, originally set up for IVIS interface design recommendations, happened to be also efficient for ADAS evaluation (see PSA experiment about ACC). |
| Drawbacks | To be careful about giving instructions to the driver regarding meanings of factors ( especially distinction between interference – using the system and driving -, and temporal demand – time constraint during the session-) |
| Limits | No detailed information about subjective evaluation of the driving task performance (the principle of this tool relied on the objective evaluation of the performance through measurement of errors for instance) |
| Minimum number of participants | No less of 12 participants |
| Modalitity of administration | Self-administered questionnaire. It is important that the tool administration happens immediately after the conclusion of the session under evaluation |
| Instructions | Written instruction reported in the questionnaire are sufficient<br>Some factors as temporal demand and interference need deeper explanations |
| Mean time to fill the questionnaire in | First administration between 3 to 7 minutes, next administrations quicker |
| Type of analysis | For less than 30 participants, the non parametric tests are recommended: Wilcoxon (to compare factors values between sessions) and Spearman (to compute correlation between factors and weighted factors values). |

**Table 9-2 – Recommendations for DALI questionnaire**

| PSA-TLX | |
|---|---|
| **Purpose** | Evaluation of ADAS or IVIS impact on driving safety (driving effort and performance) consequence on the ability to achieve the driving task<br>Evaluation of driving demands with or without any additional task (IVIS /ADAS) |
| **General description** | Driving task oriented tool specified for real road<br>Assessment of felt driving demands and management of driving activity (compromise effort and performance)<br><br>Evaluation of driving activity compared with comfortable and usual driving (familiar itinerary and vehicle)<br>- Effort (mental resources allocated to cope with task's demands),<br>- Performance (disruption, seriousness and frequency of errors),<br><br>Two levels of evaluation :<br>- Global driving activity<br>- 7 driving sub-tasks :<br>1/ Control of the lateral motion<br>2/ Control of the longitudinal motion<br>3/ Reactivity to dynamic road environment<br>4/ Reactivity to static road environment<br>5/ Itinerary following<br>6/ Use of controls and driving equipments<br>7/ Reactivity to safety and status signs<br><br>Evaluation of driver's state (fatigue, stress, discouragement, overall dissatisfaction), link to help to interpretation (no statistic correlation)<br><br>Score varies from 0 to 100 (scale resolution by 5 points) + comments/justifications<br>6 levels with meaning (no implicit and subjective meaning of scores) |
| **Advantages** | Sensitive to different types and levels of demands (driving, or driving + systems)<br>Explanations of driving effort and performance (driver's comments on scores)<br>Emphasize on driving activity management (compromise effort/performance)<br>Proposition of a decision criteria for driving effort and driving performance safety (semantic threshold)<br>explicit meaning of scores reduces risk of scores subjective interpretation<br>Lighten data analysis : most efficient analysis is compromise between effort and performance<br>Quick automatic statistical processing and graphical representation |
| **Drawbacks** | Experimenter training is required for efficient questionnaire administration<br>Too many explanations of questions<br>Some concepts are a bit difficult to understand (need clear explanations)<br>1st administration and exhaustive data analysis are time consuming<br>Changes in lower scale resolution<br>Driver's state is not linked statistically with driving effort and performance, as it is in the theoretical model the method is based on. |
| **Limits** | No targeting of peaks of workload<br>Not designed for simulator test : validation test needed<br>not dedicated for comparison of different HMI (for example two modalities of warning, or two different visual HMI for route guidance system) |

| | |
|---|---|
| **Minimum number of participants** | at least 12 |
| **Modalitity of administration** | Immediately after the driving task<br>After each test condition<br>Minimum duration of the driving task (test condition) : at least 10 minutes<br>For the first time : necessity to read the questionnaire with the driver and explain the driving sub-tasks (ensure his understanding).<br>Drivers fill in the questionnaire themselves<br>For the other times : just give the questionnaire to fill<br>Check filled questionnaire : ask for comments and justifications |
| **Instructions** | After each test condition, evaluation of driving activity in comparison to the usual and comfortable driving (subjective) |
| **Mean time to fill the questionnaire in** | Expected duration :<br>1st administration : 20 minutes<br>next administration : around 10 min |
| **Type of analysis** | Two levels of analysis :<br>1/ lighter analysis to have quick results : analysis is the compromise between effort and performance of driving, as this analysis reflects the way the driver manages the driving activity (with or without using a system) and amphasize any issue on the driving task achievement.<br>5 levels of compromise efficiency have been identified :<br>- optimum<br>- very favourable<br>- favourable<br>- unfavourable<br>- very unfavourable<br>2/ a deeper analysis more exhaustive (especially on effort and performance) is possible if time is available for deeper diagnostic |

**Table 9-3 – Recommendations for PSA-TLX questionnaire**

# 10 Innovation

As emphasized in section 2 of the deliverable, workload subjective methods currently used in most human factors studies (e.g. NASA-TLX), do not provide assistance in the design of IVIS or ADAS considering many issues : difficulty to interpret, low sensitivity… as they are not initially designed for car-driving.

In task 2.2.6, the three evaluated methods – BMDMW, DALI and PSA-TLX - are the first methods designed to evaluate mental workload for car-driving context.
Results of tests conducted within task 2.2.6 showed that all these three tools are promising attempts as they present satisfactory level of sensitivity and can be used in the design of embedded systems.

Two of these 3 tools are direct exploitation of previous work on mental workload achieved in aeronautics and in automotive research:

- PSA-TLX and DALI use the formalism already used in the NASA-TLX (multiple scales ranging from 0 to 100)
- DALI re-uses some factors already used in NASA-TLX but reject some other that are replaced by new factors adapted to automotive field evolution.
- PSA-TLX exploits research work done by Schlegel in automotive field and proposes a pragmatic tool usable by practitioners

# 11 References

Allen, T.M., Lunenfeld, H., & Alexander, G.J. (1971). Driver information needs. Highway Research Record, 366, pp. 102-115

Allard, F. (2001). Information Processing in Human Perceptual Motor Performance. University of Waterloo.

Annett, J. (2002). Subjective rating scales: science or art? /Ergonomics/, 45(14), 966-987.

Archer, R. and Lockett, J. (1997). WinCrew: A tool for analyzing performance, mental workload and function allocation among operators; in, Proceedings of ALLFN 97, Revisiting the allocation of function issues, Vol. II, pp157-165, IEA Press.

Cherri, C., Nodari, E., Toffetti, A., (2004). Deliverable 2.1.1 "Review of existing Tools and Methods". AIDE Integrated Project.

Chin, E., Nathan, F. (2004). Deliverable 2.1 "State of the art on HMI metrics and target values". RoadSense project.

Chin, E., Nathan, F. (2004). On-road evaluation with the PSA-TLX of a night vision enhancement system. PSA Peugeot-Citroën internal report.

Chin, E., Nathan, F. (2005). On-road evaluation with the PSA-TLX of a voice-recognition IVIS. PSA Peugeot-Citroën internal report.

Couyoumdjian, A., Di Nocera, F., Ferlazzo, F. (2002). Spontaneous speed: theoretical and applied considerations. In D. de Waard, K.A. Brookhuis, J. Moraal, & A. Toffetti (Eds.), Human Factors in Transportation, Communication, Health, and the Workplace (pp. 175-188). Maastricht, the Netherlands: Shaker Publishing.

De Waard, D. (1996). The measurement of driver's mental workload. Thesis ISBN 90-6807-308-7.

Deregibus, E., Rambaldini A. (CRF) (2004). D2.2 Development of test plans. Roadsense CRF Case Study, contribution to Deliverable D 2.2.
European Organisation for the Safety of Air Navigation, (2003). Review of Workload Measurement, Analysis and Interpretation Methods. CARE-Integra-TRS-130-02-WP2-1-0.

Di Nocera, F., Terenzi, M., Capponi, C., Ferlazzo, F. (2005). Behavioral Markers of Driver's Mental Workload. Cognitive Ergonomics Laboratory Technical Note n° WL01. University of Rome "La Sapienza"

Farmer, E.W., Jordan, C.S., Belyavin, A.J., Bunting, A.J., Tattersall, A.J. and Jones D.M. (1995). Dimensions of operator workload. Defence Evaluation & Research Agency, Report DRA/AS/MMI/CR95098/1.

Gautheret, L. and Nathan F. (2001). Subjective evaluation of mental workload : specification of a method for car-driving. PSA Peugeot-Citroën internal report.

Gawron, V.J., (2000). Human Performance Measures Handbook. Lawrence Erlbaum Associates.

Goillau, P.J. and Kelly, C.J. (1996) Malvern Capacity Estimate (MACE) - A proposed cognitive measure for complex systems; in, First Int. Conf. On Engineering Psychology and Cognitive Ergonomics.

Gopher, D., and Donchin, E. (1986). Workload- an examination of the concept. In, K. R. Boff, L. Kaufman and J. P. Thomas (Eds.), Handbook of Perception and Performance: Volume II, Cognitive Processes and Performance. New York, NY, USA: John Wiley and Sons. 41-1 to 41-49.

Gouweleeuw, J.M., Hughes, A.J. Mann, J.L. Odoni A.R. and Zografos, K. (1996). Global MSR studies - Part 2: Review of available techniques/facilities, MUFTIS report MFTS/DRA/WPR1.2/1.

Harper, R.P., and Cooper, G.E. (1986). Handling Qualities and Pilot Evaluation. *Journal of Guidance*, 9(5), 515-529.

Hart, S.G., & Hauser, J.R. (1987). Inflight application of three pilot workload measurement techniques. *Aviation, Space, and Environmental Medicine, 58,* 402–410.

Hart, S. G., & Staveland, L. E. (1988). Development of a multi-dimensional workload rating scale: Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload*, 139-183. Amsterdam, The Netherlands: Elsevier.

Johansson, E., Engström, J., Cherri, C., Nodari, E., Tuffetti, A., Schindhelm, R., Gelau, C. (2004). Deliverable 2.2.1 "Review of existing techniques and metrics for IVIS and ADAS assessment". AIDE Integrated Project.

Jordan, C.S. (1992). Experimental study of the effects of an instantaneous self assessment workload recorder on task performance. Defence Evaluation & Research Agency, Report No. DRA/TM(CAD5)/92011.

Lansdown, T. C., Brook-Carter, N., & Kersloot, T. (2002). Primary Task Disruption from Multiple In-Vehicle Systems. ITS Journal, 7(2), pp 151-168. ISSN: 1024-807

Leplat, J. & Cuny, (X). Introduction à la Psychologie du Travail. Paris : Presses Universitaires de France.

Luximon, A. and Goonetilleke, R. S. (2001). Simplified subjective workload assessment technique. *Ergonomics*, 44(3), 229-243.

Nathan, F. and Chin E. (2005). Workload evaluation in car driving. Presentation in National Congress of French Society of Psychology 2005, special session "Driving assistance and road safety".

Pauzie A. (1994). Human interface of in-vehicle information systems, VNIS' 94, Yokohama, Japan.

Pauzié A. & Pachiaudi G. (1997), Subjective evaluation of the mental workload in the driving context, In Traffic & Transport Psychology - Theory and Application - Pergamon, pp. 173-182.

Pearson, P., Byars, G. (1956). The Development and Validation of a Checklist Measuring Subjective Fatigue (Report No. 56-115). Randolph Air Force Base, TX: School of Aviation.

Rasmussen, J., Leplat, J., and Duncan, K. (Eds.) (1987). New Technology and Human Error, J. Wiley, London.

Reid, G.B. and Nygren, T. E. (1988). The subjective workload assessment technique: a scaling procedure for measuring mental workload. In P. A. Hancock and N. Meshkati (eds), *Human Mental Workload* (Amsterdam: North-Holland), 185 ± 218.

Roscoe, A.H. and Ellis, G.A. (1990). A subjective rating scale for assessing pilot workload in flight: A decade of practical use. RAE Technical Report TR90019.

RSSB, Rail Safety and Standards Board, (2005). Train Driver Mental Workload Acceptable Workload Evaluation Guidance Note.

Schlegel (1993). In Peacock and Karwoski, "Automotive Ergonomics", 359-382.

Tattersall, A. J. and Foord, P. S. (1996). An experimental evaluation of instantaneous self-assessment as a measure of workload. Ergonomics, 39(5), 740-748.

Terenzi, M., Capponi, C., Di Nocera, F., Ferlazzo, F. (submitted). Marcatori comportamentali del carico di lavoro mentale in compiti di guida automobilistica. Congresso Nazionale della Società Italiana di Ergonomia.

Vidulich, M. A. (1988). Notes on the AHP procedure. (Available from Dr. Michael A. Vidulich, *Human Engineering Division, AAMRL/HEG,* Wright-Patterson Air Force Base, OH 454336S73).

Vidulich, M. A.,Ward, G. F., & Schueren, J. (1991). Using the Subjective Workload Dominance (SWORD) Technique for Projective Workload Assessment. Human Factors, 22 (6), 677-691.

Wickens, C.D., and Hollands, J.G. (2000). Engineering Psychology and Human Performance (third edition). New Jersey: Prentice-Hall Inc..

Wierwille, W. W. and Casali, J. G. (1983). A validated rating scale for global mental workload measurement applications. In Proceedings of the Human Factors Society 27th Annual Meeting (pp. 129-133). Santa Monica, CA: Human Factors Society.

Wolf, J. D. (1978). Crew workload assessment: Development of a measure of operator workload (AFFDL-TR-78-165). Wright-Patterson AFB, OH: *Air Force Flight Dynamics Laboratory.*

# 12 Appendix

Appendixes jointed with D226 are excluded from the D226 report, and are independent documents.
The list of appendixes is presented below: each appendix corresponds to an independent file.

| Title | Content |
|---|---|
| Appendix I* | CRF |
| 1 | BMDMW questionnaire |
| 2 | CRF detailed results |
| 3 | PSA-TLX Compromise Effort Performance |
| Appendix II* | INRETS |
| 1 | INRETS experimental rpotocol |
| 2 | DALI questionnaire |
| 3 | Final questionnaire detailed results |
| 4 | DALI complementary results |
| 5 | PSA-TLX complementary results on effort and performance |
| 6 | PSA-TLX complementary results on driver's state |
| 7 | BMDMW complementary results |
| Appendix III. | PSA |
| 1 | PSA detailed technical report |
| 2 | PSA-TLX questionnaire |
| 3 | Procedure of PSA-TLX analysis |
| 4 | Tool for PSA-TLX data automatic processing (statistics computation) |
| 5 | Tool for PSA-TLX figure edition |
| Appendix IV. | Task 226 |
| 1 | T226 recommendations delivered to T227 |
| 2 | Final questionnaire used to evaluate the three questionnaires – BMDMW, DALI and PSA-TLX. |

* only one file is available for all appendixes.