Large Scale Collaborative Project

7th Framework Programme

INFSO-ICT 224067

D.2.2.2 Testing and Evaluation Strategy II

Deliverable n.	D2.2.2	Testing and Evaluation Strategy II	
Sub Project	SP 2	FOT Framework	
Work package	WP 2.2 (as the base line)	Methods and tools	
Task n.	Several in SP2, SP3, and SP4		
Author(s)	Franzén, S. Karlsson, I.C.M. Paglé, K. Morris, A.	File name	TeleFOT_D2.2.2_Testing and Evaluation Strategy II_130222.doc
Status	Final		
Distribution	Public (PU)		
Issue date	2013-02-22	Creation date	2012-11-30
Project start and duration	1st of June, 2008 – 54 months		

European Commission
Information Society and Media

Project co-funded by the European Commission

(DG-Information Society and Media)
DG-Communications Networks,
Content and Technology

in the 7th Framework Programme

SEVENTH FRAMEWORK PROGRAMME

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ABBREVIATION	DESCRIPTION
CAA	Cockpit Activity Assessment Module
CAN	Controller Area Network
D	Deliverable
DAS	Data Acquisition System
DoW	Description of Work
D-FOT	Detailed FOT
GDS	Green Driving Support
GPS	Global Positioning System
GPRS	General Packet Radio Service
GSM	Global System for Mobile Communication
FOT	Field Operational Test
L-FOT	Large Scale FOT
NAV	Navigation System
PI	Performance Indicator
SA	Speed Alert
SP	Sub Project
TI	Traffic Information
IR	Internal Report
WP	Work Package

REVISION CHART AND HISTORY LOG

REV	DATE	AUTHOR	REASON
1.0	2012-10-31	Stig Franzén MariAnne Karlsson Katia Paglé Andrew Morris	Synopsis
2.0	2012-12-26	Stig Franzén	First draft
3.0	2013-01-07	Stig Franzén	Second draft
3.1	2013-01-15	Marianne Karlsson	Comments on second draft, incl. additional material
4.0	2013-01-16	Stig Franzén	Final draft submitted
4.1	2013-02-22	Stig Franzén	The final version completed taking the reviewers' comments into account
4.1	2013-03-01	Petri Mononen	Formatting edits to finalise

EXECUTIVE SUMMARY

This deliverable, *D2.2.2 Testing and Evaluation Strategy II,* provides the verified and validated description of the recommended *general strategy (or overall methodology)* that was applied in and across the different FOTs in TeleFOT.

The general strategy includes the overall, recommended approaches for generating research questions and hypotheses, for choosing study design, for recruiting and choosing participants, as well as a description of what type of data is to be collected and recommendations for data collection methods and procedures to be used.

Some assumptions, definitions, and other relevant facts are listed here:

- A Field Operational Test (FOT) is defined as a test that is run under normal operating conditions in the environment typically encountered by the subjects and the equipment being tested.
- A FOT involves a larger number of users using the services and systems in their daily life in actual use conditions.
- The TeleFOT project system boundary is the larger context of traffic, transport and travel
- The users are found in related roles, i.e. as drivers, passengers and travellers
- The platform used for the functions was not a vehicle but aftermarket and nomadic devices, i.e. a wider approach to the work was taken;
- Pre-trip, during trip, and post-trip use of devices and functions needed to be investigated rather than in-vehicle use only;
- The impact areas stretched well beyond safety issues and included also efficiency, environment, and mobility,
- User uptake (user acceptance and adoption) was addressed in its own right (user uptake is not an impact area)
- The combination of a top-down and a bottom-up approach to generate research questions and hypotheses provides an important approach to FOT evaluations in general (also outside the transportation sector).

1 INTRODUCTION

The general objectives of the TeleFOT project was firstly, to assess the impacts of functions provided by aftermarket and nomadic devices used in vehicles for driver support and secondly, to raise the awareness of the functions and potential that these devices offer.

The assessment of impacts was achieved by running a series of Field Operational Tests, or FOTs, at different locations across Europe. These local FOTs tested different systems and functions ranging from traffic information, navigation support, and speed alert to green driving support. It was acknowledged that the preconditions of the different FOTs varied. However, in order to increase the possibility for generalisation, and to accomplish an assessment of the impacts of the different systems and functions on a European level, a common evaluation regime and generic approach – a strategy – was deemed necessary.

1.1 Objective and scope

The deliverable, D2.2.1 *Testing and Evaluation Strategy I,* reported on the preliminary work accomplished in WP2.2 Methods and Tools during the first year of the TeleFOT project. The main objective of WP2.2 was to define and develop evaluation plans for the test schemes identified for each test community. The work was to continue during the whole project.

This deliverable describes the knowledge gained at the point of the project completion in November 2012. It should be seen as a recommendation for a verified and validated *overall methodology* to be applied in and across different FOTs. It addresses the study regimes in terms of procedures for generating research questions and hypotheses, for choosing study design, for recruiting and choosing participants, and for data to be collected and for data collection methods and procedures.

During the last three years of the TeleFOT project the material in this deliverable constituted input into the work of SP3 and SP4 and the work carried out in SP3 and SP4 constituted an important feedback for the update of the Testing and Evaluation Strategy. The strategy (described in this Deliverable D2.2.2 Testing and Evaluation Strategy II) was developed and operationalized in dialogue between the relevant partners involved in SP2, SP3 and SP4.

It should be noted that reference should now be made to the latest version (5.0) of the FESTA Handbook. The TeleFOT contributions to the revision of the FESTA Handbook (in the context of the project FOT-NeT) have been reported in Deliverable D2.1.2.

One major change compared to the preliminary version of the strategy is the emphasis on the importance of subjective measures and how they can be incorporated in an assessment work and that also user uptake can be included. Another is the importance of the data management process as a whole from the very beginning of an FOT activity. This was manifested in the TeleFOT project as a "Data Working Group" where contributions from SP2, SP3 and SP4 partners played an important role throughout the project.

However, still some basic facts are valid: (i) that a number of different devices and functions were tested and in different combinations; (ii) that the platform used for the functions was not a vehicle but aftermarket and nomadic devices, i.e. a wider approach to the work was taken; (iii) that pre-trip, during trip, and post-trip use of devices and functions needed to be investigated rather than in-vehicle use only; (iv) that the impact areas stretched well beyond safety issues and included also efficiency, environment, and mobility, (v) that user uptake (user acceptance and adoption) was addressed in its own right (user uptake is not an impact area), and furthermore (vi) the impact of the tested functions and devices on the transport system as a whole.

1.2 Accomplishment

The work of D2.2.1 (i.e. the preliminary strategy) was carried out as part of TeleFOT WP2.2. as a collaboration between the main partners (i.e. CHALMERS, CIDAUT, CRF, IKA,

LOUGHBOROUGH, and VTT). The common approach was discussed and decided upon by the partners, and the partners then had the main responsibility for compiling information on different matters whereas all had the responsibility for commenting and contributing to the deliverable as a whole. The work was completed by the end of the first project year.

This deliverable (D2.2.2) is the result of a joint effort from the SP2, SP3 and SP4 leaders with Chalmers as the lead partner. It is based on the structure of D2.2.1 (as much of the first approaches proposed have been working well) but with amendments made based on material and findings from work in SP2, SP3 and SP4 relevant for the enhancement and validation of the strategy.

1.3 Structure of deliverable

The deliverable is structured as follows:

- a description of the key characteristics of TeleFOT with implications for the study design etc. (Chapter 2)
- a description of the process for generating and choosing hypothesis (Chapter 3) including a tentative list of common research questions and hypotheses;
- a description of the process for choosing relevant performance indicators (Chapter 4);
- a description of the principles for choosing study design (Chapter 5);
- a description of the principles for the selection of participants in the tests (Chapter 6);
- a description of how to deal with issues concerning test environment (Chapter 7);
- a description of the principles for the type of data that should be collected and the methods by which to collect the information (Chapter 8).
- a description of test procedures, pilot tests etc. (Chapter 9);
- a description of the ethical and legal issues to be considered (Chapter 10).
- Chapter 11 provides a guideline for the implementation of the strategy and a description of the approach by which the strategy should be implemented across FOT
- a concluding chapter (Chapter 12) summarising the results.

2 GENERAL APPROACH

2.1 The key characteristics of TeleFOT FOTs

TeleFOT is an example of an FOT, or rather several FOTs. In order to define the key characteristics of the TeleFOT project and its different FOTs, it has been necessary to first describe the following key concepts: field test, naturalistic study, experiment, and field operational test.

- By *field tests* is generally meant to test something, e.g. a product, under actual operating conditions or in actual situations reflecting intended use.
- By *naturalistic study* is most often referred to a study where researchers/corresponding observe and record some behaviour or phenomenon, often over a longer period of time, in its natural setting while interfering as little as possible with the subjects/participants or the phenomena.
- *Experiment* is defined as a test or trial carried out for the purpose of discovering something unknown or of testing a principle, supposition, etc.
- By *controlled experiment* is meant an experiment that isolates the effect of one variable on a system by holding constant all variables but the one under observation.
- A *Field Operational Test* is generally described as a test run under normal operating conditions in the environment typically encountered by the subjects and the equipment being tested. Normally a FOT involves a larger number of users using the services and systems in their daily life in actual use conditions.

The TeleFOT project consists of Large scale FOTs (L-FOTs) and Detailed FOTs (D-FOTs). L-FOTs are *naturalistic studies* in the sense that they are studies in which was investigated normal, everyday use of a set of different functions provided by the platforms of nomadic and aftermarket devices. The studies concerned conditions in which the participants received, used and reacted to functions and services provided to them and data was collected over a longer period of time from a larger number of participants. The studies were also *experiments* in the sense that tests were undertaken in order to

find out the answers to questions and hypotheses posed. Nevertheless, they were not controlled experiments even though a rigid test procedure was executed. It should be noted that L-FOTs can provide knowledge about a function in use without necessarily make use of a hypothesis testing procedure as there are research questions that are not fitted for such an approach.

Also the D-FOTs were carried out as experiments in the sense that the tests were undertaken in order to find out the answers to research questions and hypotheses posed. However, even though not all D-FOTs were not carried out as completely controlled experiments, the D-FOTs were run with more control than the L-FOTs, e.g. the participants were asked to drive certain routes, as well as under certain conditions. Furthermore, less vehicles and less participants were involved and the vehicles were equipped with additional equipment why more, as well as more detailed, data could be collected (e.g. on acceleration patterns, speed, petrol consumption, etc.) than was the case in L-FOTs.

The L-FOTs constituted the core of the TeleFOT project. The main purpose of running L-FOTs and D-FOTs across different test sites, and in parallel, was to benefit from the particular strengths of the respective approaches, i.e. the higher ecological validity[1] of the naturalistic driving test results, providing evidence of behaviours and behavioural changes over time, and the higher reliability of the controlled experiment, offering possibilities to identify more causal types of explanations to e.g. the drivers' behaviour. Thus, the D-FOTs are a complement to the L-FOTs, providing further information for the support of the analysis and interpretation of the results.

2.2 Project boundaries

[1] Ecological validity concerns to which degree the experimental findings mirror what one can observe in the real world (ecology= science of interaction between organism and its environment). Ecological validity is whether the results can be applied to real life situations and not only to an experimental setting.

Field Operational Tests, FOTs, have been carried out in Europe, in Japan, and in the U.S. Even though the general objective of these types of tests can be described as testing new transport technologies in general, a common focus of previous projects has been in-car use of driver assistance systems, in particular different safety systems.

The purpose of TeleFOT was to assess the impacts of other functions provided by different aftermarket and nomadic devices. As stated in the Description of Work (DoW) the project was to test "… driver support functions…", but also that "….mobile services may have impacts on the whole travel chain – not only on driving". In contrast to several of the former FOTs, with mostly a narrow focus and a system boundary defined by driver–vehicle–road, the *TeleFOT project system boundary is the larger context of traffic, transport and travel, and in addition the user in different roles, i.e. as driver, passenger and traveller*, respectively.

The project further addressed the use of functions and services in relation to pre-trip (e.g. the use of functions for planning a trip), during trip (e.g. the use of functions to re-plan a trip or to help way-finding), as well as post-trip (the use of a function, e.g. green-driving support, to assess the outcome of a trip). This means that data had to be collected beyond that of in-vehicle use and there was eminent a necessity to collect information also by means of travel diaries and questionnaires in addition to the data gathered by different logging devices.

Furthermore, earlier FOTs have often had a focus on safety related to driving. Also in TeleFOT one of the impact areas concerned driver behaviour and safety but other areas were also considered. These include: efficiency, environment and mobility, as well as socio-economic impacts on the transport system as a whole.

In addition was addressed user uptake[2] (i.e. issues related to acceptance and adoption of new technology and new behaviours) which can be considered a prerequisites for the former mentioned impacts. In TeleFOT, all L-FOTs addressed all impact areas whereas D-

[2] Usability is frequently mentioned in the DoW. User Uptake is the wider concept including aspects related to e.g. the usability of the devices and the functions.

FOTs had a more narrow focus. For instance was mobility considered difficult to address in a D-FOT.

3 RESEARCH QUESTIONS AND HYPOTHESES

3.1 General guidelines

Some general guidelines for the formulation of research questions and hypotheses apply to the TeleFOT project as a whole, and to the different FOTs. However not ALL research questions of interest may result in a hypothesis. Some research questions can be addressed without the use of hypothesis testing procedures.

3.1.1 Formulating hypotheses

A hypothesis is a statistically testable statement of an anticipated effect that is a refinement of a more general research question. For example, an FOT may be designed to answer the following research question "*How does the use of a lateral warning system affect safety?"* The resulting hypothesis may be termed "*Drivers who use a lateral warning system exhibit safer driving behaviour*".

In formulating a hypothesis, consideration should be given to the variables under scrutiny. It is vital that the variables collected in an FOT allow acceptance or rejection of the hypotheses. To do this, both the independent and dependent variables should be well defined at the start of the FOT. The independent variable is one which can be manipulated by the researcher[3]. As the researcher changes the independent variable, he or she records what happens using dependent variable(s). The resulting value of the dependent variable is caused by and depends on the value of the independent variable. Other variables, known as controlled or constant variables are those which a researcher wants to remain constant and thus should observe them as carefully as the dependent variables. Most studies have more than one controlled variable.

The hypothesis should be stated in the following typical formats:

[3] Researcher should here be understood as the person(s) involved in planning and running the different FOTs.

H1 – Younger drivers will be more likely to find smart phones easy to use than older drivers

or

H2 – Drivers will travel less distance during the course of a week when using a navigation system compared to when one is not used.

Each of these hypotheses states the measure (activation of the system or conflict situations), the direction (more likely, fewer) and the comparison conditions (young vs. older, engaged vs. disengaged).

Within TeleFOT, both these formats were used in order to address the impacts in the different impact areas. For example, H1 above relates to issues concerned with user uptake and the differences that may be apparent among a population of users, whereas H2 is an exposure issue that relates to safety and environmental impacts when the system is in use.

3.1.2 Prioritizing hypotheses

Within the context of TeleFOT, it was important that the number of hypotheses to be tested was kept to the minimum whilst ensuring that all impact areas as well as user uptake were covered. This helped focus on those hypotheses where a real effect was expected that would have a significant impact in each of the respective impact area. An approach to prioritizing the hypotheses is presented in section 3.2.3.

3.1.3 From hypothesis formulation to concluding on results

The studies must be designed to enable the research questions and hypotheses to be tested. In order to test the example H1 above, for instance, a sufficient numbers of drivers in each of the two demographic groups would need to be recruited in order to be able to make a robust comparison. For H2, the researchers should be able to define the appropriate measures to record in advance of the data collection.

The techniques used in the data analysis should be appropriate to the hypothesis under consideration. For H1, a *between–groups analysis* would be the best to be undertaken on

the frequency rates of activation of the system (comparing Older vs. Younger drivers). For H2, a *within–groups analysis* would need to be undertaken to compare the two conditions (System ON vs. System OFF).

The results section of a FOT report should make direct reference to the hypotheses, as stated in the experimental design. The report should state whether the hypothesis was supported or not. Furthermore, the concluding section of the FOT report should, where appropriate, highlight those hypotheses which were supported, and those which were not. In the case of the latter, a possible explanation for the findings and suggestions for further avenues of research should be given. In the case of the hypothesis being supported, conclusions should be tentative, and caveats noted, particularly when the findings are novel.

3.2 Generating hypotheses top-down and bottom-up

According to the first version of the FESTA Handbook the formulation of use cases provided the main basis for generating research questions and hypotheses to be tested in a FOT. However, this approach was insufficient for the TeleFOT project with its many functions and combinations of functions and given that all functions were not determined at the beginning of the project.

As part of the work on the strategy for the TeleFOT project, a generic modified process for generating hypotheses for different FOTs was developed. For the TeleFOT project an integrated Top-down and Bottom-up approach was used to firstly generate the research questions and subsequently formulate the hypotheses. This integrated method ensured that the hypotheses had a foundation in both the theoretical aspects of the impact area under consideration (Top-Down) and in the system functionality (Bottom-Up considering both intended and unintended effects). This in turn helped ensure that all the potential impacts were identified. These approaches are explained in more detail in the following sections and are summarised in Figure 3.1.

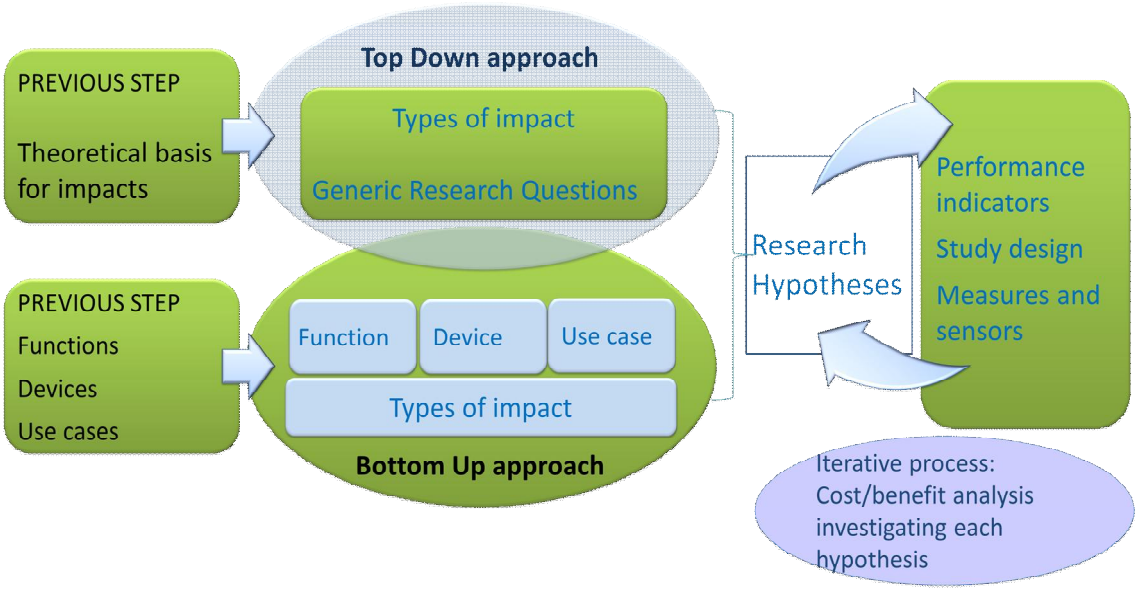

Figure 3.1 A combined top-down (impact area) and bottom-up approach for the generation of research questions and related hypotheses

For TeleFOT, the generic research questions and hypotheses generated using this top-down approach for the impact areas efficiency, environment, mobility and safety are available in other documents.

3.2.1 The top-down approach

The basic principle for generating hypotheses using a top-down approach lies in a theoretical understanding of the factors that influence the different impact areas. It should be noted that there is likely to be overlaps of these factors among the impact areas under consideration and hence the same research question and resulting hypotheses will be applicable across more than one impact area. The approach will result in *generic research questions* that are independent of the any system functionality. It can be applied also in other FOTs and for other, additional devices and functions than those initially tested in TeleFOT.

The procedure for generating hypotheses in a top-down approach (Figure 3.2) is as follows;

- The impact area should be considered in its entire context and primary measures affecting that area be identified.
- Secondary factors of these measures are then identified that can be used to explain the variations in the primary measures.
- Finally the variables affecting the secondary measures are identified.
- The variables identified form the basis of the generic research questions ''*Is there a change in the variable?*'' and the hypothesis based upon an anticipated effect of the variable ''*The variable will increase/decrease.*''
- This leads to the generic hypotheses that can be tested in a statistical manner. The direction each hypothesis should take (e.g. increase or decrease) is based upon the anticipated effect once the top-down approach is integrated with the bottom-up (system defined) approach.
  - *Journey lengths will increase/decrease when the system is used compared to when it is not used.*
  - *Journey duration will increase/decrease when the system is used compared to when it is not used.*
  - *The number of journeys will increase/decrease when the system is used compared to when it is not used.*
  - *The use of rural roads/motorways/major roads will increase/decrease when the system is used compared to when it is not used.*

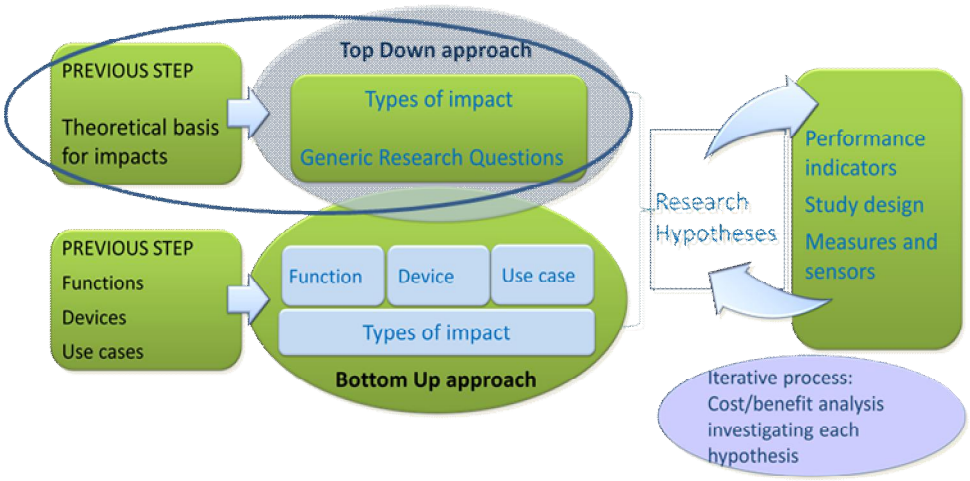

Figure 3.2 Top Down Approach

This procedure should be undertaken for each of the impact areas, i.e. for Efficiency, Environment, Mobility and Safety.

3.2.1.1 Efficiency

Traffic efficiency is usually defined as the extent to which a certain transportation supply can meet the travel demand of people in a transportation system. Efficiency is perceived differently by road authorities and road users. In TeleFOT efficiency is discussed from the point of view of the road users.

Traffic efficiency level of the network was estimated based on indicators such as speed, travel time, headway, etc. In the following figure (Figure 3.3) the primary measures are Traffic Flow, Traffic Volume, and Accessibility. The secondary measures are for Traffic Flow; Travel Time, Delays, Speed, and Time/Distance Headway. For Traffic Volume it is Time/Distance Headway, and for Accessibility they are Traffic Jams and Delays.

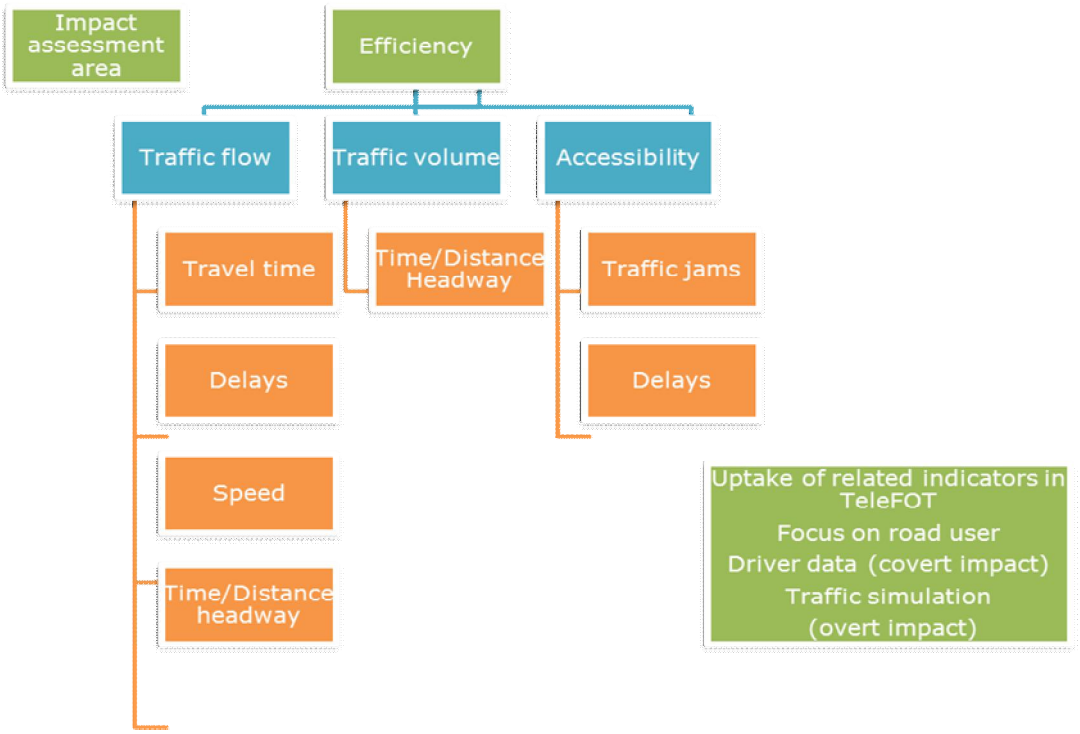

Figure 3.3 The TeleFOT Efficiency Model

3.2.1.2 Environment

The environment impacts have the basis in the following primary measures; Fuel consumption, CO_2 emissions, Route choice, Transport mode, Average speed, and Speed distribution.

3.2.1.3 Mobility

Mobility is defined as the potential for movement. It consists of the means of transport and the transport networks to which one has access, knows about and is willing to use. Mobility is the willingness to move together with potential and realised movement rather than just physical movements of vehicles, people and goods. Typically frequent journeys should be studied (e.g. commuting trips) and the impacts can be found on different dimensions as indicated in the following figure (Figure 3.4), where the primary and secondary measures are highlighted.

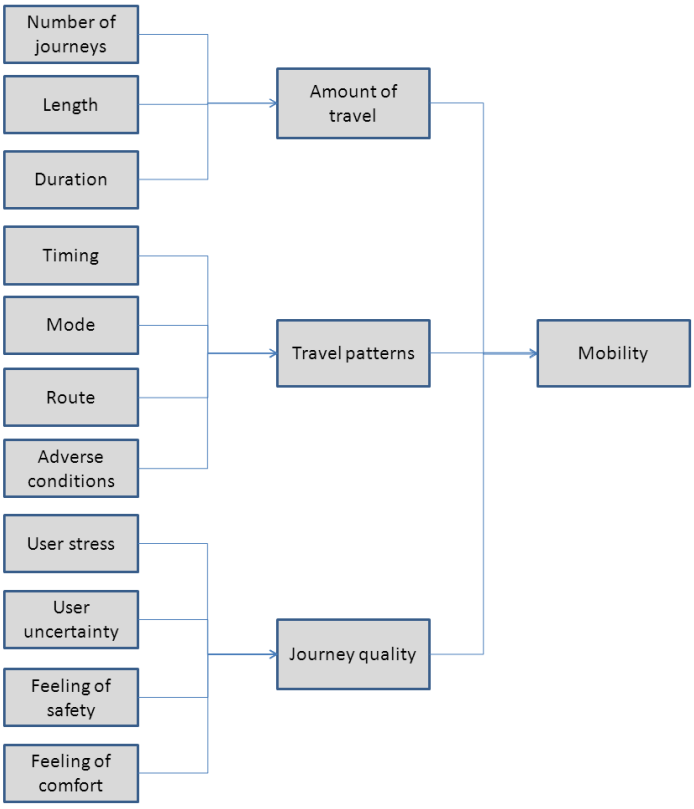

Figure 3.4 The TeleFOT Mobility Model

3.2.1.4 Safety

The primary measures affecting safety would be the ‘Number of Events (accidents, near misses) that occur’ and the ‘Severity of the Event’. Secondary factors affecting the first of these measures would, for example be ‘Exposure of the vehicle on the road’, ‘The driving style of the driver’, ‘The distraction of the driver from the driving task’ and ‘Any interaction with the fitted device’. Considering the factor ‘Exposure’, this can be measured with the following variables: ‘Length of journey’, ‘Number of trips undertaken’ and ‘Road type used’. These variable lead research questions addressing Route choice, Journey length and duration, Average speed and speed violations, Lane positioning, Braking behaviour, “Hands off wheel” time, and Visual distraction.

3.2.1.5 User Uptake

User Uptake is not an impact area. It should rather be considered as a prerequisite for the assumed impacts. However, the same procedure regarding primary and secondary measures can be used. In Figure 3.5 the relation between the primary and the secondary measures for User Uptake are indicated. User Uptake is defined as the extent to which users adopt and integrate the nomadic devices and the functions offered into their everyday life, i.e. invest in them, use them and make use of the functions in relation to planning and undertaking journeys by car and/or by other means of transport. Acceptance is seen as the extent to which users approve of the nomadic devices and functions tested and are willing to keep them for future usage.

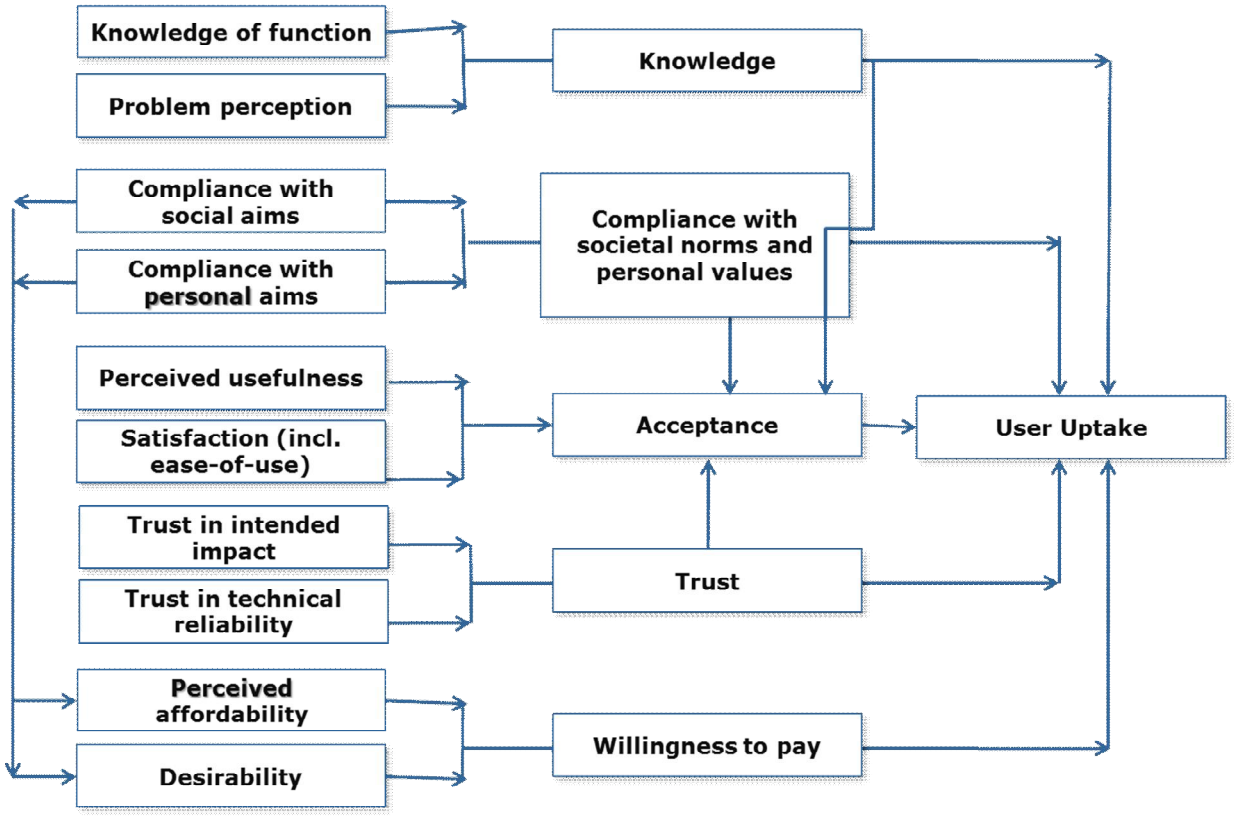

Figure 3.5 The TeleFOT User Uptake Model

3.2.2 A bottom-up approach

The bottom-up approach can only be undertaken once the functions, devices and use cases have been defined for the particular test site. For each impact assessment, the detailed research questions should be developed based on a consideration of four sets of factors.

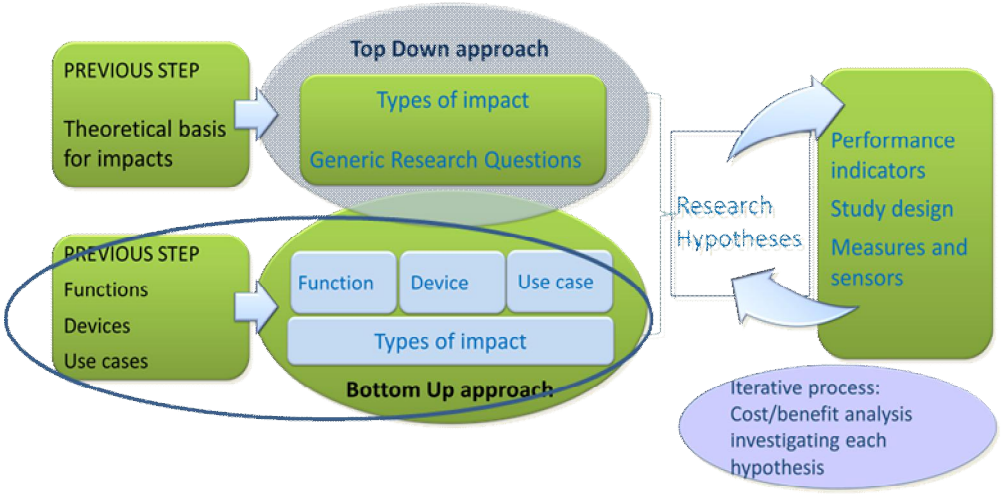

Figure 3.6 Bottom Up approach

1. Function: A functional description of the system (i.e. what it does) and the effect that the function may have on a user in the context of each Assessment domain.
2. Design: The implementation of the system (i.e. how it is designed), and the impact the design attributes have on the user–system interaction in the context of each Assessment domain
3. Use Case: The use cases (i.e. the context of use factors) and their relationship with consequences of use within the real world
4. Types of Impact: The types of Impacts that are being considered.

The following are examples of hypotheses that are based on a generic Navigation Support system. For the sake of the example it is assumed that this system can be installed and used with and without a visual display depending on the driver's choice. The hypotheses are aimed to address the *Safety Impact*.

Hypothesis 1: *There will be fewer instances of sudden braking with the device in use.*
Reasoning behind hypotheses based upon the system: *Drivers will receive information in time to control vehicle speed on the approach to manoeuvres.*
The safety implications are: *There will be a reduction in rear-end collisions.*

Hypothesis 2: *There will be a change in the allocation of visual attention when the device is in use.*
Reasoning behind hypotheses based upon the system: *A new (visual) information source has been introduced that will be diverting visual attention from the road in the lead up to making a manoeuvre.*
The safety implications are: *A change in the number of incidents/accidents due to visual inattention*.

Hypothesis 3: *Personal road miles will increase when the device is used.*
Reasoning behind hypotheses based upon the system: *There will be increased confidence (and therefore willingness) to undertake unfamiliar journeys by car rather than choosing an alternative transport mode.*
The safety implications are: *The accident rate will increase due to an increased exposure in terms of miles driven.*

For TeleFOT, a tentative list of hypotheses according to the bottom-up approach has been generated and is reported elsewhere.

3.2.3 Integrating and choosing hypotheses

In this stage of the process a huge number of research questions and associated hypotheses from the top-down and the bottom-up approaches have been developed. A key task was to integrate both sets of hypotheses in the context of each FOT. It was envisaged that the bottom-up approach would form the basis of the hypotheses list for a FOT and that the top-down approach would be used in order to check that nothing significant for a particular impact area was omitted.

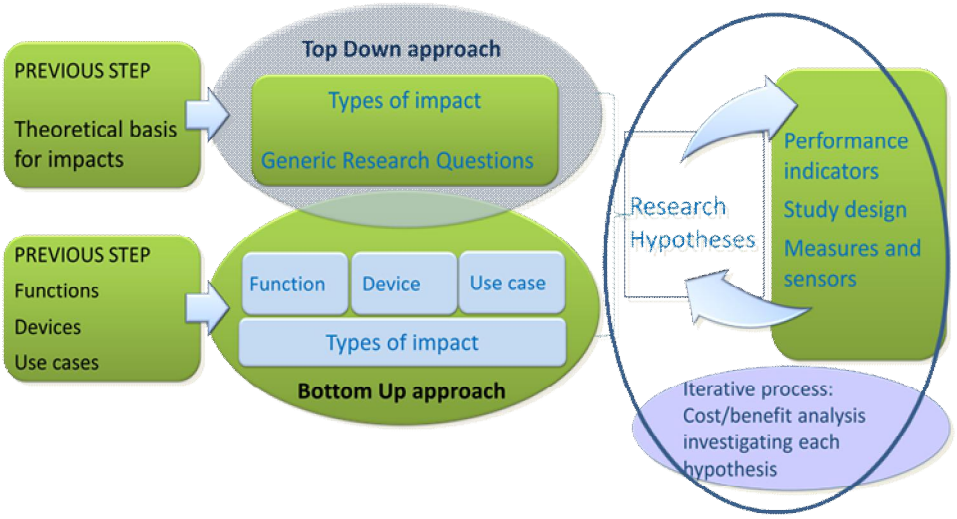

Figure 3.7 Choice of hypotheses – the cost-benefit approach

After the integration had taken place, the list of hypotheses was still very long. In order to derive a final, manageable set of research questions and hypotheses that could be applied throughout the various test sites, a cost–benefit approach was proposed and performed (Figure 3.7). Using this approach, an assessment was made regarding the likely "costs" of collecting the data. Costs were represented in terms of the effort required to derive a performance indicator expressed predominantly in terms of resources. This should be offset against the likely "benefit" that proving/disproving the hypotheses would have. This was measured by way of the likely contribution towards providing a significant answer the research question and thus the level of contribution to the impact assessment. To some degree, this would depend upon the stakeholder needs and requirements, and therefore a prioritisation of their 'needs' was considered.

The following table (Table 3.1.) illustrates one example the way in which the cost benefit approach could be applied to one specific Hypothesis H3: *Eyes off road time will be increased when using a SatNav (Navigation Support function) compared to when one is not used*. It can be seen that the strategy has three parts that are used to determine the costs and the benefits.

- Part one states the hypotheses and hence defines the minimum required performance indicators (PIs) to evaluate the hypotheses. The collection of these PIs is a cost.
- Part two considers the influencing factors and explanation for the result seen in the hypotheses. This requires additional data that can be explored. This additional data is also a cost.
- Part three considers the relative contribution that the hypotheses under consideration will make towards understanding the assessment of the impact in each area. This is a benefit.

If the cost of collecting all the required data (e.g. video footage, microphone for audio levels, eye tracking equipment etc.) beyond the basic data acquisition system or GPS does not reflect in the anticipated benefit in terms of assessing the impact, then the hypotheses should not be included in the analysis plan.

Table 3.1. Strategy for prioritising hypotheses (Example: Safety Impact related to a Navigation Support function)

Part	What it states	Example	Why it is needed	Example
1	The hypothesis (stated for a defined function, device, use case) (Should also consider 'when' e.g. initial use and after prolonged use?)	Eyes off road time will be increased when using a SatNav with a visual display & auditory/visual distance countdown where there are several turnings within a short distance compared to when one is not used.	It determines the Performance Indicators (PIs) that are required (in detailed or large scale trials)	Number & length of glances away from the road scene.
2	The reasoning (behind the hypothesis)	Because where there are multiple choices, the driver will not be able to rely only on the auditory information to be sure of the manoeuvre to take and will frequently look at the visual information.	It determines the other data that must be collected to fully understand and explain the findings. (in detailed or large scale trials)	Number & length of glances to the display. Visual & auditory information being presented by system. Location of vehicle. Road layout. Driver statement explaining their experience in using

				the system.
3	The implication (for safety)	The risk of inattention to an incident in the road scene will be increased.	It identifies the importance of testing the hypothesis	Priority of testing this hypothesis would be based on number of accidents that are caused by inattention in such scenarios.

3.3 Tentative list of common research questions and hypotheses

A common set of research questions and hypotheses were to be addressed in the TeleFOT FOTs. A tentative list of common research questions and hypotheses was generated based upon the top-down and bottom-up approaches, and by taking into account their feasibility in D-FOTs and L-FOTs (Table 3.2.). The list had to be further reduced.

However, from the tentative list can be concluded that one specific research question may result in several different hypotheses depending upon the function(s) to be tested, thus the top-down approach must be complemented by the bottom-up approach as described in the former section. It can also be concluded that one and the same hypothesis can, given the anticipated impact of the function to be tested, have a positive or a negative direction. The approach followed made these issues apparent. Furthermore, it can be assumed that not all aspects would have been covered in such a systematic way, across all functions, without the top-down approach.

The list of hypotheses generated on basis of the top-down approach was used for generating hypotheses for all functions tested in TeleFOT hereby ensuring a consistency across the different FOTs of TeleFOT.

Table 3.2. A tentative list of common research questions and hypotheses, generated on basis of a top-down and bottom-up approach.

* Apply to almost all hypotheses and is therefore not repeated. (GDS=Green Driving Support, NAV=Navigation Support, SA=Speed Alert/Speed Limit information, TI=Traffic Information)

Research question	Hypothesis	Functions			
		GDS	NAV	SA	TI
Is the travel time(s) from origin to destination affected?	Travel times will increase due to lower speeds (with access to function compared to without).*			x	
	Travel times will decrease because the system provides information on fastest routes.		x		
	Travel time will decrease for regular trips because the driver chooses smoother routes or timing based on the information provided.				x
	Time spent travelling will increase because speed is decreased due to traffic related warnings.				x
Are any delays avoided?	There will be less delays because the driver is informed about incidents causing delays.				x
	There will be less delays because the system supports the driver to find a more efficient route.		x		
Are any traffic jams avoided?	There will be less exposure to traffic jams because the driver chooses an alternative route to avoid the incident(s).				x
	There will be less exposure to traffic jams because the system reroutes the driver to other routes in case of such an incident en route.		x		

Is speed affected? Is average speed affected?	There will be a decrease/increase in speed because the driver is warned about incidents ahead. (The highest speeds are cut). The variation in speed will increase because drivers react differently to the information. There will be a decrease in speed because the driver is warned about exceeding the speed limit. Speed limit violations will decrease because the system reminds/warns the driver. The variance in speed will decrease because the driver obeys speeds limit better.			x x x	x
Is speed homogeneity affected?	Speed homogeneity will increase.	(x)			x
Are speeds of single vehicles more harmonized?	Drivers will choose more constant speed in order to avoid extra fuel consumption (or to minimize consumption)	x			
Is there a change in target speed?	There will be a decrease in target speed to be upheld because of access to information on …				x
Is acceleration affected? Is very sudden/ heavy acceleration affected?	There will be a decrease in incidents of sudden acceleration/deceleration. There will be a decrease in harsh/sudden braking because the driver anticipates the situation due to warnings.	x			(x) x
Is gear choice /RPM affected?	Higher gears will be used to a higher extent which leads to lower rpm. (and hence …)	x			
Is the total amount of fuel used affected?	There will be an increase in total (average) fuel consumption if		(x)	x	

Is average fuel consumption affected?	roads with higher speed limits are chosen (motorways). There will be a decrease in total (average) fuel consumption if roads with lower speed limits are chosen (rural roads, highways).		(x)	x	
Is the emission of CO_2 affected?	There will be an increase/decrease in CO_2 emissions (depending on whether the consequence is less time spent/trip or more journeys are undertaken).		x		
Is the number of journeys undertaken affected?	The number of journeys undertaken will decrease as the driver becomes more aware of environmental impacts. The number of journeys will increase/decrease because it will be easier to travel and find places also in unfamiliar environments.	x	 x		
Is the duration (in time) of journeys affected (for regular trips)?	There will be longer travel times as a consequence of lower (average) speeds. There will be a decrease in time spent travelling because the driver will choose a smoother route (or time) based on the information provided. There will be a decrease in time spent travelling because less time is spent trying to find the route or place.		 x	x	(x) x
Is the distances travelled affected (for regular trips)?	There is likely to be an increase in distances travelled because the driver is warned about incidents ahead (and chooses an alternative, longer route) There will be a decrease in distances travelled because of a reduction in detours.		 x		x
Is there a change in time allocated to travel (for regular trips)?	More time will be allocated to travelling.	x		x	
Is there a change in travelling in adverse conditions (dark, fog, slippery road, etc.)?	There will be an increase in journeys undertaken in dark conditions because the driver is guided by the system.		x		(x)

Is there a change in route choice (for regular trips)?	There will be a change in route choice for regular trips because of information on incidents ahead. Drivers will increase their use of roads with higher speed limits (motorways). There will be an increase in 'rat running' because in some cases the system reroutes the driver to minor roads.			x	X x
Is there a change in travel mode (for regular trips)?	There will be an increase in the use of private cars as the increased access of information leads to increased travel comfort. There will be a decrease in the use of private cars as the driver becomes aware of the many disturbances in the traffic. There will be a decrease in the use of private cars as the drivers get more aware of the (negative) environmental impacts.	x			X x
Is there a change in time (of day) for travelling (regular trips)?	Travellers will start earlier or postpone their regular journeys because of access to information on incidents.				x
Is 'eyes off road time' affected (frequency, duration)?	'Eyes off road' (time and frequency) will increase because the driver allocates visual capacity to read the information. (Refers to solutions where the information is provided in a textual format).	x	x	x	x
Is mental workload affected?	There will be a decrease in mental workload because the driver has access to information on …		x	x	
Is there a change in users' level of stress?	There will be a decrease in user's level of stress because of better access to information on … There will be an increase in stress level due to annoying, repetitive messages	x	x x	x x	 x
Is there a change in users' uncertainty?	There will be a decrease in the users' experiencing uncertainty as		x	x	

	the user is informed about …				
Is there a change in users' feeling of travel comfort?	There is an increase in users' estimation of travel comfort as the user is better informed about … There will be an decrease in travel comfort due to annoying, repetitive messages		x x	x x	(x) x
Is there a change in users' feeling of subjective safety?	There is an increase in safety because the driver experiences driving more carefully (according to speed limits) There is an increase in safety because the driver experiences driving without disturbances (accidents, incidents, queues)		x		(x)
Have the (perceived) mobility options increased?	The (perceived) mobility options will increase in the following ways: …..		x		(x)
Is there a change in perceived journey quality?	There will be an increase in perceived journey quality because of a perceived feeling of control There will be a decrease in perceived journey quality because of a decrease in personal integrity	x	x	x	x
Is there a change in users' knowledge?	The users' knowledge of the function will increase over time (and increase the potential for user uptake)	x	x	x	x
Is there a change in users' problem perception?	Users' perception of problems associated with environmental issues/negative environmental impacts will increase. Users' perception of problems associated with mobility issues will increase. Users' perception of problems associated with safety (speeding) will increase.	x		 x	 x

Is the device/function considered useful?	The higher the perception of usefulness, the more substantial the user uptake.	x	x	x	x
Is the device/function considered satisfying?	The higher the perception of satisfaction, the more substantial the user uptake.	x	x	x	x
Does the user trust the device/function?	The higher the perception of trust, the more substantial the user uptake.	x	x	x	X
Is the user willing to invest in the device/function?	The higher the perception of usefulness, satisfaction, trust etc., the more likely the user will be to invest in the device/function.	x	x	x	x

3.4 Hypotheses testing in D-FOTs versus L-FOTs

The final formulation and choice of research questions and consequent hypotheses provided the basis for the design of tests and evaluations.

Within the TeleFOT project data was gathered in both L-FOTs and D-FOTs. The manner in which these tests were undertaken differ (Figure 3.8); whilst the aim of L-FOTs were to collect core data over a longer period of time in, as far as possible, naturalistic driving conditions, the D-FOTs were carried out under controlled experimental conditions.

These differences had implications for the nature of the hypotheses posed and the way in which they could be answered. It should be mentioned that the idea behind a D-FOT was that they should address topics related to a L-FOT set-up that could not be addressed by other means than what a D-FOT set-up could offer. They are complementary to the L-FOT concerned.

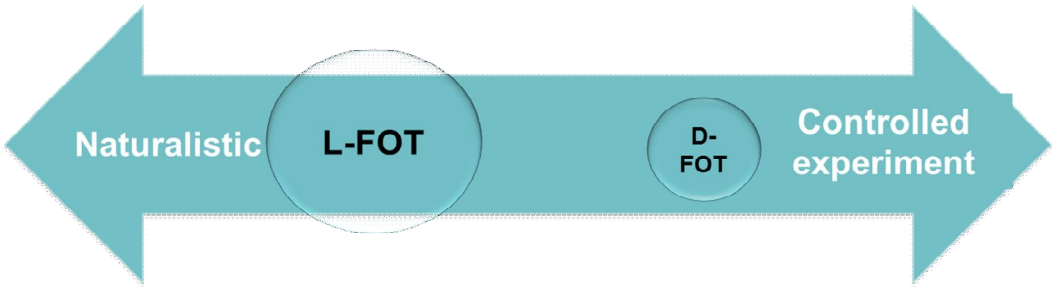

Figure 3.8 L-FOTs vs. D-FOTs

For the L-FOTs it was apparent that apart from the data collection from data loggers (incl. GPS signals from devices, etc.) much of the data collected would be subjective data, gathered primarily from questionnaires and travel diaries. Rigorous statistical testing of subjective data is challenging since the data may not be in the form for which hypothesis testing is intended. Particular care is required when interpreting the results since much underlying variability, which is impossible to control, may be more influential than the presence or absence of the device. In addition, if a Between Subjects design is

chosen then subjective questions of preference, either with or without the device, become extremely difficult to answer.

For the D-FOTs, there was the opportunity to collect more of logged, objective data that could directly measure aspects of a device was in use, e.g. related to HMI issues. With careful experimental design that accounts for many confounding factors and variability, the D-FOTs should also in some specific cases be able to prove/disprove hypotheses by way of statistically robust and rigorous analyses.

4 PERFORMANCE INDICATORS AND MEASUREMENTS

4.1 From hypothesis to measurement

After defining the research questions and hypotheses, appropriate performance indicators (PIs) must be chosen to answer them. Performance indicators are by definition quantitative or qualitative measurements, expressed as a percentage, index, rate or other value, which is monitored at regular or irregular intervals during the test and which can be compared with one or more criteria.

A denominator (per time interval/per distance/in certain location/…) is needed for each indicator as this makes the measures comparable. Indicators are to be agreed on before the tests are run (but it may be possible to develop a few new or combinations of indicators during the course of the study). Indicators should cover all impact areas of TeleFOT. Furthermore, they should be validated so that evidence of the related impact area is available, e.g. based on previous research.

The performance indicators will be defined based on research questions and hypotheses. For each research question and hypothesis at least one indicator should be provided. However, the same indicator may be appropriate for several hypotheses. Typically, one research question leads to several hypotheses – in this phase the set of statements expands. For a specific test site the number of indicators and measurements will probably be less than the number of hypotheses (Figure 4.1).

The schematic picture (Figure 4.1) indicates the process for one impact area but the connections (arrows in the figure) may extend from one impact area to another. For example, an indicator such as ‘journey length’ will be valid for the impact area ‘safety’ since it gives a measure of exposure on the road but will also be valid for the impact area ‘environment’ since increases/decreases in journey lengths relate to emissions. It is important to be systematic in this endeavour in order to cover all impacts/impact areas.

Measurable indicators should be defined for all hypotheses considered relevant for a function or combination of functions. The discussion on top-down and bottom-up processes (cf. Chapter 3) and use cases should be completed on the level of hypotheses and research questions. For practical reasons, some indicators may finally be left out. It should be acknowledged that this will have an impact on the hypotheses and research questions.

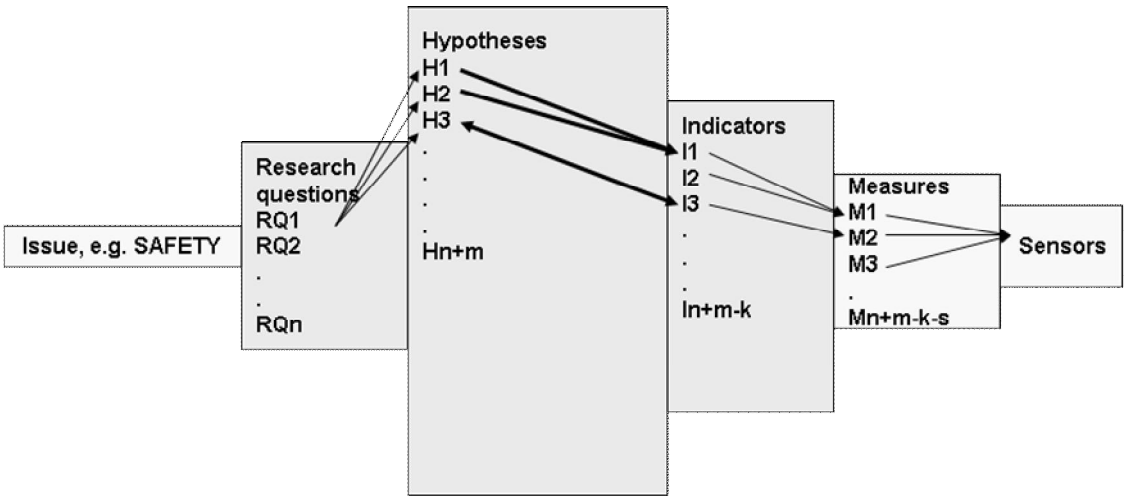

Figure 4.1 Description of the procedure – From impact area and research questions to measures and sensors.

The FESTA indicator table should be utilized as a source of information when defining the indicators. The FESTA table includes indicators for different impact areas and indicates how reliable and valid the suggested indicators are. For example, for speed changes there are altogether 10 indicators. The hypothesis in question defines which of these should be chosen.

The FESTA table also serves a checklist for how well the hypotheses cover the research questions. The selection of measurements and sensors combined to the indicators are shown in the FESTA table. In the following tables (Tables 4.1-4.3), three examples indicate the path from research questions and hypotheses to indicators.

Table 4.1 The link between research question and measure given the function Traffic Information and a message in road conditions (e.g. danger of aqua planning).

Research question	Hypothesis	Indicators	Measure/sensor
Is speed affected?	The driving speed is reduced due to the warning.	Mean speed m/s or kph	Speed GPS

Table 4.2 The link between research question and measure given the function Traffic Information and the message “Road works 25 km ahead on E18”.

Research question	Hypothesis	Indicators	Measure/sensor
Is road type affected?	The exposure is shifted from higher road types to lower because the driver wants to avoid the road works.	Km motorway / km rural road	Road type GPS

Table 4.3 The link between research question and measure regarding the participant’s level of intention to use the system.

Research question	Hypothesis	Indicators	Measure/sensor
Is the system in use (on/off)?	The driver keeps the system on.	Behavioural intention.	Verbal report/rating Questionnaire

5 STUDY DESIGN

Each FOT should be designed so as to enable the chosen hypotheses to be tested in a statistically rigorous manner. Since many of the hypotheses aim to show the effect the introduction of a 'treatment' (e.g. the use of a Navigation Support System) it is essential that *base-line* or *control data* is collected along with the experimental data so that comparisons can be made between user response when, for example, the Navigation Support System is in use compared to when it is not.

There are two general study designs that can be used to generate data in both control and experimental conditions; Within Subject Design and Between Subject Design.

5.1 Within Subject Design

In a Within Subject Design, each participant undertakes a period of time in both the control condition and the experimental condition. For example in a FOT on driving without and with a Navigation Support System, every participant drives for some time without (control condition) and for some time with the navigation support system (experimental condition).

This type of design has two main *advantages*:

(i) fewer participants are needed than in the case of a Between Subject Design, and

(ii) it is more likely to find a significant effect, given the effects are real since there is no variance in the characteristics of the subjects in the experimental and control groups.

However, there are also *disadvantages.* There is a risk for so called carry-over effects, which means that if a participant experiences one condition, this may affect his/her driving in the other condition. For example, in a D-FOT where the control and

experimental data are both collected for a pre-determined route the first measurements taken may influence the second since people gain familiar with a specific road. Furthermore, there should be a sufficient amount of time (and therefore exposure) allocated to the control condition and the experimental condition in order to experience similar circumstances in both conditions (e.g. weather, lighting, traffic density etc.) but also allowing for adequate experimental data to be collected.

Typical sequences in an FOT would be O_x X_x or O_y X_y O_y where O represents baseline condition and X represents the experimental (or treatment) condition (Figures 5.1 and 5.2).

Figure 5.1. A study design with a control and an experimental phase (O_x X_x).

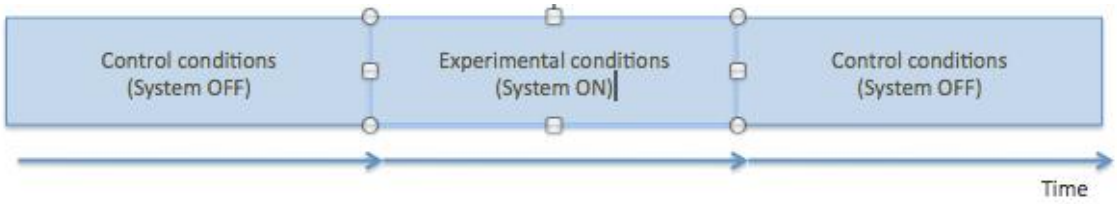

Figure 5.2 A study design with control phase, experimental phase, and control phase (O_x X_x O_x).

The superscripts x and y numbers indicate, for instance, the number of weeks or months that the sequence will last (cf. Popkin et al., 2003) or the distance to be travelled (cf. Regan et al., 2006). In the D-FOTs, the participants' exposure to control or experimental conditions will be fairly short. In the L-FOTs the exposure should be considerably longer in order for e.g. new behaviours to be established. Furthermore the control conditions

and the experimental conditions should preferably be of equal length, e.g. 6 months of control conditions (system OFF) and 6 months of experimental conditions (system ON)[4].

In case the number of devices/loggers available is small the subjects in the within subject design can be split into two groups. The first will undertake the control condition followed by the experimental condition and the second the experimental condition followed by the control condition (Figure 5.3.).

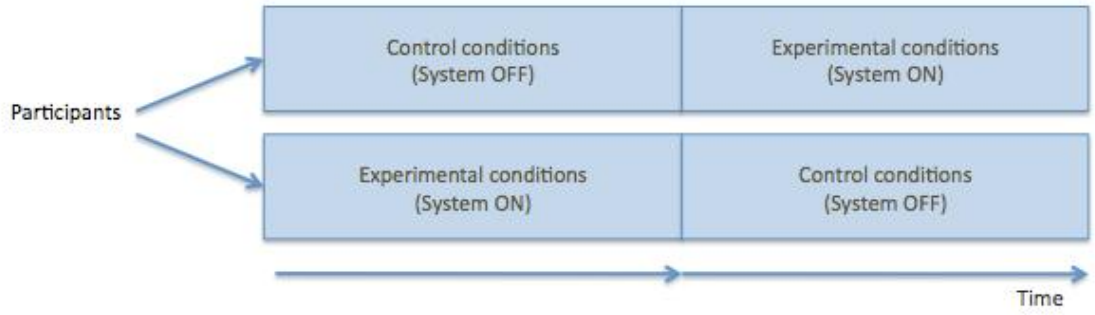

Figure 5.3 A study design with control phase and experimental phase. The participants are split in two groups and randomly assigned to groups A or B.

5.2 Between Subject Design

In a Between Subject Design each participant takes part in either the control condition (control group) or the experimental condition (experimental group) (Figure 5.4.). The groups should then be equal in size. Thus, for the same amount of data, twice as many

[4] However, in several earlier FOTs, corresponding to the L-FOTs carried out in TeleFOT, baseline conditions have lasted for fairly short periods of time (2-3 weeks), and treatment conditions for a slightly longer but still fairly short period of time (between 3-4 and 8-9 weeks, i.e. between 1-2 months). When distance travelled have been used as a way to determine exposure, baseline conditions may have prevailed for 150-300 km and treatment conditions for 300 km (or more) (cf. Regan et al. 2006). A 3-4 weeks treatment period may suffice but it may also be too short depending upon the overall purpose of the test (and the hypothesis to be tested) and the total number of participants involved.

participants are required in a between subject design than for a within subject design. The number increases further if more than one treatment is under consideration.

The advantage of a between subject design is that carry over effects are not a problem as individuals are measured only once in every condition; one measurement is completely independent of the other. Furthermore, the duration of the FOT is shorter for a between subject design than a within subject design since both the control group and the experimental group can run in parallel.

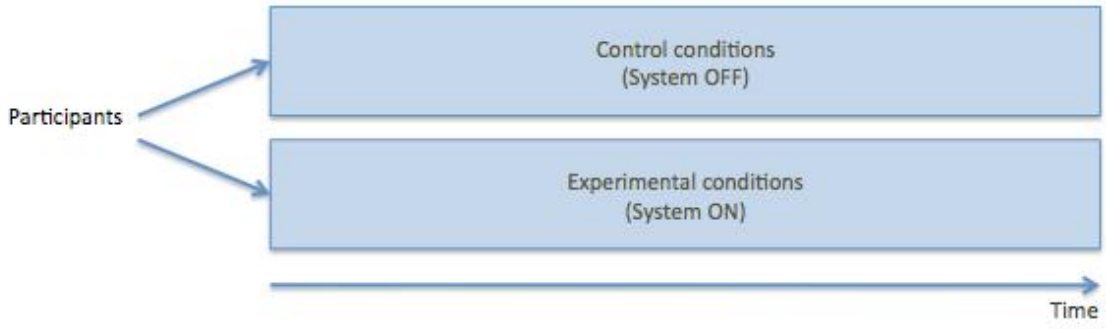

Figure 5.4. A study design where control conditions and experimental conditions are achieved by a Between Subject Design.

However, there is the strong possibility of confounding effects due to differences in the characteristics of participants in the control group and the experimental group. In order to limit these effects a *random assignment* of participants between the groups can be undertaken. Alternatively can the groups be matched on pre-selected characteristics, e.g. age, gender, driving experience and/or attitude towards new technology. In order to do this it is necessary to identify in advance the variables that need to be matched. By creating two groups of matched pairs, matched on all those factors that are considered to be of importance, the extraneous factors, which may cause confounding problems, will be controlled to some extent. The main drawback with the matched pairs design is the sampling process. As the number of characteristics that requires matching increases, a correspondingly large sample pool will be required to allow adequate matching. A further problem is that this design assumes that the researcher actually knows what extraneous factors need to be controlled for, i.e. matched - and in some circumstances this may not be the case.

5.3 Choosing between study designs

Considering the pros and cons of the Within Subject Design and the Between Subject Design it is preferential in the TeleFOT context to conduct a Within Subject Design as this, though more time consuming, requires fewer subjects and is more likely to produce significant analysis results. If a Within Subject Design is not possible then the Between Subject Design should be employed with careful consideration given to matching the control and the experimental groups.

5.4 Dealing with integrated functions

The designs described above assume that a single function is under observation in the FOT. If it is the case that several and/or integrated functions are being tested (and this is the case in several of TeleFOT FOTs) then further consideration needs to be given to the study design. This applies to the 'experimental condition' in the within subject design as well as the 'experimental group' in the between subjects design.

If the (integrated) functions can be isolated (switched on and off independently of each other) then more complicated study designs can be implemented that will test the impact of each individual function. However, if the functions cannot be isolated then they must be considered as a single treatment and any conclusions from the FOT can only be made about the combined effect of the functions being applied simultaneously.

Further information relating to more complicated study designs for integrated functions, assuming they can be isolated, can be found on the FOT-NeT webpage (www.fot-net.eu)

6 PARTICIPANTS

6.1 Randomised and non-randomised sampling

A random sampling procedure is often considered the most desirable method to select the participants for a study. In the case of FOTs, this is not feasible and other, non-randomised, methods must be used. Nevertheless, a random procedure could be used in order to assign participants to the corresponding experimental condition (control group or experimental group) specified by the particular experimental design in each FOT (see Chapter 5).

A fundamental criterion for choosing participants is that the participants reflect the intended user population of the tested function/system. A characterisation of the intended user population is therefore a first step in the choice of participants.

A second criterion is that the participants allow for the chosen hypotheses to be tested. A stratified sampling method may be applied. Depending upon the research questions and the functions to be tested, there could be a need to select certain groups of participants taking into account some particular characteristics in such a way that these variables will act as covariates in the analysis in order to study differences between groups. Thus, the variables to be considered should define exhaustive and mutually exclusive groups, i.e. a participant should be assigned to a group and to one group only. Moreover, it is recommended to define each group of a proportional size to the corresponding strata of the total population under study in order to be able to adequately represent it.

6.2 Type of participants

As previously mentioned, the participants in the FOTs should be representative of the intended user population of the systems and functions tested. The intended user population could be described on the basis of different characteristics, e.g.:

- *demographical variables* such as age, gender, socio-economic status, etc.;
- *experience*, such as driving experience (total kilometres driven, kilometres during last year, years of driving license, usual types of roads driven, etc.), experience of technology (embedded systems, nomadic and after-market devices, etc.), and experience of functions (navigation support, speed alert, etc.); or
- *personality and attitudes*, considering aspects such as sensation seeking, locus of control and/or attitudes towards road safety issues (new technology, public transport, etc.).

However, on the one hand and depending upon the chosen research questions, there may be a need to select a particular group of participants for inclusion in the FOTs, ensuring that this group is in some way representative of those users who will ultimately interact with the system. On the other hand, considering too many factors in the choice of participants will increase the requirement for the number of participants.

Based on the use case descriptions, the basic criteria for the selection of participants in TeleFOT FOTs are:

- (as a general approach) the participants should be between 25 and 65 years old. Nevertheless, each FOT should select the age range of the participants considering the age allocation of the driving population within each country;
- the participants should have more than 3 years driving experience; and
- the participants should drive more than 10 000 km/year.

Before beginning the recruitment for any of the FOTs, the researchers/corresponding must consider the relationship between individual differences and the behaviour which the function/system is seeking to influence. The reason for this is that recruiting on a personality or attitude basis ensures that a system is tested on a broad range of drivers who may interact with the system very differently, given that this kind of variables are very likely to have a direct influence on behaviour. This way, it would be possible to explain differences in driver behaviour and system use.

As a consequence, depending on the systems and functions to be evaluated in TeleFOT (i.e. particular functions tested), certain specific individual differences (e.g. driver's

attitudes towards technology) would need to be considered within the sampling process, setting up a stratification method, as explained earlier.

6.3 Number of participants

FOTs are described as studies in which a large number of individuals participate. In TeleFOT a difference has been made between D-FOTs and L-FOTs with implications for the number of participants to be involved. A test involving e.g. 10-12 participants will not be regarded as a L-FOT whereas a test involving, e.g. 100 participants may. On the other hand, 10-12 participants may suffice in a D-FOT.

The required number of participants in an FOT will always depend upon a number of factors, e.g. the number of functions and/or systems to be tested, the hypotheses formulated, and the choice of a between or a within subjects design, etc. If the number of participants is small, it is difficult to statistically prove any effects of the function/system that are actually there whereas a large number of participants increase the chance of finding an effect. On the other hand, a large sample implies a higher investment in terms of equipment, resources, etc.

In order to ensure that the chosen sample size is representative for the behaviour of a group of users and that it is possible to statistically prove any effects that are there, a power analysis is needed to calculate the desirable sample size. A statistical power analysis exploits the relationships between the four variables involved in statistical inference: sample size, significance criterion, population effect size, and statistical power (e.g. Cohen, 1992). For any statistical model, these relationships are such that each of them is a function of the other three. In line with the previous reference, the following formula can be adopted as a basic approach to determine the sample size to be used in the trials:

$$n = \frac{Z^2 p q}{E^2}$$

where n is the size of the sample, Z is the confidence level (determining the confidence of the generalisation of the data from the sample to the population), p is the estimated effect, q is equal to 1-p (where pq represents the level of variability in the computations to verify the hypothesis) and E error (the percentage of error to be accepted in the generalisation).

If considering e.g. a confidence level of 95% (Z=1,96), an estimated effect of 5% (p=0,05 and q=0,95) and an allowed error of approx. 2,5% (E=0,025), the required number of participants would be n=292. Nevertheless, this final number would correspond to the total number of observations needed. Thus, depending on the registered data:

- if only a single observation per subject is stored, 292 subjects would be needed.
- if more than one observation per subject is registered (and this will be the case for most of the variables in L-FOTs as well as D-FOTs, e.g. will a large amount of speed values be stored): less subjects would be needed since for each of them, a large set of will be available.

This should be considered as a first estimation of the appropriate number of subjects/observations to be considered in a L-FOT since effect sizes, acceptable errors, etc. cannot be assured in advance.

6.4 Drop outs

During an FOT, in particular an L-FOTs, one has to take into account that not all participants will finalise their participation as planned. It is important to consider that such *drop-outs* can have a biasing effect on the results of the FOT if the participants who quit early differ systematically from those who finalise as planned with regard to relevant characteristics (e.g. socioeconomic status, age, gender etc.).

For this reason, it is necessary to anticipate and plan for participant 'drop-outs' throughout the FOTs. There are two options. One is that the selected sample is larger than required already from the beginning of the FOT (in relation to the selected

experimental design). However, this option may cause extra costs which may not be feasible. A second option is to create a “replacement pool”, i.e. one is to ‘reserve’ an extra number of participants to be recalled at short notice.

7 STUDY ENVIRONMENT

7.1 Environmental conditions

The study environment is a critical element in all FOTs, since it will determine the data that is collected and the ability to fulfil the objectives and test the hypotheses. Such environmental conditions include:

- geographical location, e.g. types of roads, traffic patterns, infrastructure and communication issues, transportation options;
- weather conditions, e.g. sun, fog, rain, snow;
- time of day; and
- seasonal effects, including e.g. the seasonal opening/closure of roads

The fundamental idea behind the TeleFOT project is that the functions and devices to be evaluated should be used in everyday life and during as natural use conditions as possible. Therefore, designing an L-FOT that *requests* the participants to e.g. drive certain roads or during certain weather conditions contradicts the overall idea of FOTs, which is the real-life usage of functions and services. Research questions and hypotheses that require very specific environmental conditions in order to be tested, e.g. very particular weather conditions or very particular types of roads, should be consequently avoided in the L-FOTs whereas they may be applicable in the D-FOTs.

However, if the research questions and hypotheses chosen require specific environmental conditions in order to be tested also in L-FOTs, it is recommended this should be ensured by other means. The design of the study may still increase the likelihood of the desired environmental conditions to occur. One such study design issue is the choice of participants. One could choose participants according to driving patterns, i.e. choose participants who often drive long distances on motorways as opposed to participants who drive predominantly in the city centre or choose participants according to geographical location, i.e. including participants who live in city centres as opposed to participants who live in rural areas. Also the choice of time of year is a study design issue, i.e. choosing to

run the L-FOT during the winter months as opposed to the summer months (only) will increase the probability of slippery and icy roads.

Another feasible approach is to use post-trial categorisation of key variables to enable exploration of data, e.g. an analysis of the effect of a particular function during short as opposed to longer trips or during sunny as opposed to rainy weather conditions. This requires, however, that relevant issues are identified beforehand, and that data is captured to enable matching with hypotheses and scenarios of interest.

8 DATA COLLECTION AND MANAGEMENT

In TeleFOT, a first common set of data is to be collected across test sites and FOTs in order to ensure a cross-site analysis. With this approach the European dimension of the project can be addressed and the results of the separate FOTs can be compared across the different test sites. In addition, a second set of common data will be collected across those L-FOTs and D-FOTs that evaluate the same type of functions and/or devices while a third set of data consists of data that is specific for each site. This will be collected per site only.

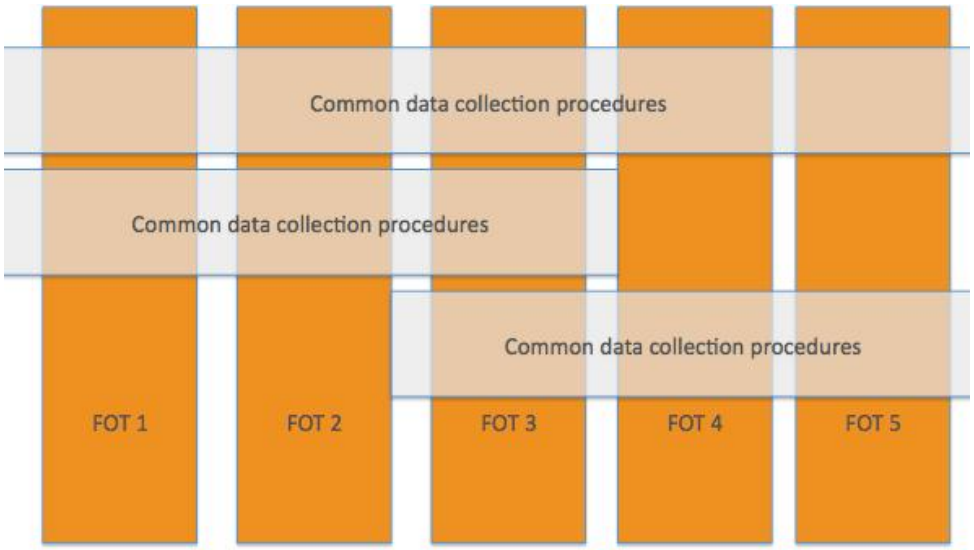

Figure 8.1. The overall approach for data collection across the different FOTs

Overall the data collected should be kept as limited as possible, and be restricted to a minimum. At the same time the data have to be sufficient for the evaluation of the defined hypotheses, both common and specific to the functions and test sites.

8.1 Data collection phases

In general, data collection is to be carried out at three points of times: pre-test, during test, and post-test.

8.1.1 Pre-test

Pre-test data will be collected in advance to the tests. This data contains information about the background of the test participants, their experiences of transportation (and driving in particular), and their attitudes towards the tested functions/systems as well as their experience of different nomadic devices and different functions/services.

8.1.2 During test

The main part of the data will be collected during the test, i.e. during the control/baseline conditions and during the experimental conditions. The data to be collected includes objective data, as well as subjective data (see Section 8.2) by means of several different types of methods and tools (see Section 8.3).

Whether L-FOT or D-FOT, the objective data will primarily be collected by means of different logging devices in the vehicles for the entire trip (or trial) and for every trip (or trial) that is undertaken in the vehicle. However, as indicated earlier, the objective data collected in the D-FOTs will be considerable more, and more detailed, than will be the case in the L-FOTs.

Also regarding the subjective data to be collected during the trials, there will be differences between L-FOTs and D-FOTs. In the L-FOTs subjective data will be collected at certain intervals and for certain periods of time. This includes data on the participant's attitude towards the function(s) and the device(s) which is to be collected: (i) after a very short period of testing; (ii) after some months of testing; and (iii) towards the end of the trial). In the D-FOTs the same information will (in most cases) be collected once the individual trial is completed.

In the L-FOTs data will also be collected by means of travel diaries. These will address the use of the different devices and the functions pre-trip, during trip, and post trip. The diaries will be distributed at set intervals during the baseline/control as well as during the experiment conditions and are to be filled in by the participants for every journey carried out during the course of one week. This should be accomplished at least once during baseline conditions (preferably more depending in the duration of the condition) and then at repeated intervals during the execution time of the FOT, e.g. every second or third month during the experimental conditions (if a within subject design) and by control group as well as experimental group (if a between subject design). The period keeping a diary should be set in a normal working period and not in holiday seasons.

8.1.3 Post-test

Data should also be collected subsequent to the different tests, in L-FOTs and in particular in D-FOTs. Post-test data collection concerns primarily subjective data in terms of information about the experiences that the test participant has had when using the nomadic device and its different functions during the trials. Herein information about the influence of the system on e.g. the participant's perceived driving behaviour, driving style and on his/her subjective judgement of the systems can be obtained. Also the participant's inclination to invest in different devices and function should be collected at this time.

8.2 Data to be collected

In order to ensure that the research questions are answered and the chosen hypotheses can be tested, subjective data as well as objective data have to be collected in both L-FOTs and D-FOTs. Also the user uptake studies will rely heavily on subjective data collection. By *subjective data* is meant data which is modified or affected by personal views, experience, or background. By *objective data* is meant data which concern facts or conditions without distortion by personal feelings, prejudices, or interpretations. In addition, *demographic data* on the participants has to be collected in order so support the interpretation of the results.

8.2.1 Demographic data

Data that should be acquired before the trials from all participants include background information on:

- age;
- gender;
- profession;
- education;
- year since obtaining a driver's license;
- distances driven per annum;
- type of road(s) basically used;
- type of car(s);
- type of car use; and
- former experience of nomadic devices and different functions.

Background data should also cover if the participant completes commuter, or very frequent journeys as well as wherefrom and whereto.

This data is not related to performance indicators but will provide an important basis for the analysis and interpretation of the results of the different FOTs, both L-FOTs and D-FOTs.

8.2.2 Subjective data

Also subjective data will be related to several performance indicators for the evaluation of the functions and the devices.

In addition to the demographic data, the data to be collected before the trials include the respective test participant's:

- self-assessment of his/her driving style;
- safety perception;
- anticipation regarding the impact of the functions/devices to be tested; and
- attitude towards new technology in general and the functions/devices to be tested in particular

The type of data that should be collected during and after the trials include:

- the way the participants have perceived the usability of the system(s);
- the participants' attitudes towards and acceptance of the devices/functions (i.e. related to user uptake);
- the perceived impact of the system/function on
    - driver behaviour
    - efficiency
    - environment
    - mobility
    - safety

In addition should be collected information on the reasons behind the different responses, e.g. whether or not the system behaved as anticipated and if errors occurred etc.

8.2.3 Objective data

A large amount of different performance indicators will determined based on the vehicle data. Therefore the data to be logged has to be matched with the function and the hypothesis to be evaluated, so that the required data is available in the evaluation phase of the TeleFOT project. For instance if the hypothesis proposes an impact on safety by reduction of speed, speed has to be logged during the test. However, the feasibility to collect data will also have to be considered in choosing the research questions and hypotheses to be tested.

8.2.3.1 Logged data

A first kind of data to be logged concern *environmental conditions*, like the vehicle position, type of road used, time and date of use, and weather conditions. In addition to the environmental conditions the inboard conditions, e.g. the usage of functions/systems or use patterns of the driver should be logged. A difference has to be made between the L-FOTs and the D-FOTs due to the limited access to data in the vehicles involved in the L-FOTs in comparison to the instrumented vehicles of the D-FOTs. In the D-FOTs the environmental conditions will be determined by logs of the GPS position and time, by

which the vehicle's position, the type of road used, and the date and time of use can be determined. Furthermore the outside temperature and the wiper status can be logged to determine the weather conditions.

For the evaluation of the drivers' activities and attention (a typical D-FOT task using an instrumented vehicle) a Cockpit Activity Assessment Module (CAA) could be used for recording the eyes-of-road time and the gaze direction. The activities can be logged by the nomadic device, which is registering all key presses and interactions between driver and device. In the L-FOTs, on the other hand, the GPS position and time should be logged to determine the type of roads, and the date and time of usage.

A second type of data to be logged is concerned with *vehicle status*. These data can describe the behaviour of the vehicle and could contain, e.g. vehicle speed, lateral and longitudinal acceleration, yaw angle, steering wheel angle, throttle position, use of brake pedal, distance towards preceding vehicle and position in lane.

For the evaluation it will be necessary also to *authenticate the individual driver/participant*. In the L-FOTs it can be anticipated that the vehicle will be used by more than one user. The authentication can be done in different ways, e.g. as an interrogation via a device in which the driver has to fill in the needed data, by a coded key, or in case the test vehicle is equipped with a driver monitoring camera, by photo shoots of the driver at the start of each trip. If none of the above options are possible, and this will most probably be the case in L-FOTs (even though not desired), the driver of the car must be documented by a separate 'log' kept by the participant him-/herself.

Information about e.g. the status of the system (on/off) or the functions used during the test runs should be logged to assure a correct evaluation of the data. However, pre-trip and post-trip use of different devices and functions will in most cases not be feasible to log. This type of data must be collected by means of other data collections methods, primarily by means of travel diaries.

8.2.3.2 Derived measures

Objective data can be linked to several performance indicators that can be read from the data set directly but a number of performance indicators can also be derived from the data during the evaluation process. Examples of such derived performance indicators are:

- To objectively judge the impact of a system on the driver, details about the driving style, e.g. the average, minimum, maximum, median and the standard deviation of the vehicle speed can be determined and used.
- A matching of the speed data with data of a digital map can deliver the number of speed limit violations and the maximum exceeding of the speed limit. (As Speed Limit Information and Speed Alert are two of the functions to be tested, this possibility is a must.)
- With access to vehicle speed and the distance towards preceding vehicles the time headway and the time to collision can be derived. For these values the mean, median, maximum and minimum can also be evaluated. In addition, both vehicle parameters can be used to determine the time spent in traffic congestions.
- With the help of map matching the road type chosen during the trips, the trip length and distribution as well as parking time and place can be determined. (Several of the common hypotheses concern choice of road types, trip lengths etc.)
- The driven distance and the fuel consumption result in consumption per distance. With access to vehicle data this information can be used to determine the pollutant emissions, such as carbon dioxide, carbon monoxide, methane, non-methane hydrocarbons, nitrogen oxides and carbon black. (As environment is one of the impact areas to be investigated, distance and fuel consumption are data that should be collected.)

8.2.4 Additional data to be collected

Additional data to be collected from the participants in the L-FOTs include:

- travel patterns;
- the use of different devices and functions pre-trip, during trip, and post-trip; and
- the perceived effects of this usage.

In the D-FOTs no travel diaries will be issued, as these journeys will not have the character of 'normal travelling' but are undertaken only for research purposes. The estimation of perceived effects of using different functions will instead be addressed in post-test questionnaires (or interviews).

8.3 Data collection methods and tools

In both D-FOTs and L-FOTs data will be collected by means of different data collection methods. Overall, a methodological triangulation approach will be applied in TeleFOT. By triangulation is meant the application and combination of several research methodologies in the study of the same phenomenon. Methodological triangulation involves using more than one method and may consist of within-method or between-method strategies.

As already stated, data to be collected involve subjective as well as objective data. Subjective data can be retrieved from, e.g. participants' descriptions of events and/or assessments of and preferences for features. Objective data can be retrieved from, e.g. recording of participants' driving behaviour and the logging of speed, acceleration, etc. as described in the former sections, but also from observations.

8.3.1 Collection of subjective data

Data collection methods that could be used for the collection of subjective data include questionnaires, individual interviews, and focus group interviews.

8.3.1.1 Questionnaires

A questionnaire is a research instrument consisting of a series of questions for the purpose of gathering information from respondents. Most often the questions, their arrangement, as well as their answers are determined in advanced. The particular benefits of questionnaires are that they are an efficient way to gather data from a large number of respondents, and that the same study can be repeated several times over across a longer time period. Furthermore, a questionnaire that is used effectively can gather information on e.g. the overall performance of a system/function tested as well as information on specific components of the system/function. The questionnaire can also

include demographic questions which can be used to correlate performance and satisfaction with the function among different groups of users. However, questionnaires normally imply a loss of control over the data collection situation which may e.g. result in a loss of data. Another disadvantage is that questionnaires does not allow for follow-up questions and probing procedures which are important in order to penetrate e.g. the underlying reasons for a certain assessment of choice. Thus, questionnaires are tools primarily for collecting data on 'what' while questions on 'why' most often need to be addressed by other data collection methods, such as personal or group interviews.

In TeleFOT, questionnaires were used in both L-FOTs and in D-FOTs. In L-FOTs, where a large number of participants were involved, questionnaires provided the main data collection method for collecting this type of information. However, it was recommended that also other methods, i.e. individual interviews and or focus group interviews, were used in both L-FOTs and D-FOTs in order to complement and further develop on e.g. the reasons behind use/non-use or acceptance/rejection. In the L-FOTs, these more in-depth question based data collection methods were used in order to elicit information from a defined sample of the larger group of test participants.

The questionnaires were used for collecting data on test participants, i.e. on their background, their driving behaviour, driving experience, and previous experience of different functions and devices (compare Section 8.2). Questionnaires were also used for monitoring the participants' perception and assessment of e.g. the contribution of the functions on the different impact areas efficiency, environment, mobility, and safety. In addition, questionnaires were used to address user uptake issues, including the participants' use of function, acceptance of function, and willingness to pay (compare Section 8.2). Changes in perception and attitudes over time were a particular issue why the questionnaires had to be distributed to the participants during baseline conditions, and during initial as well as during later phases of the experiment (compare Section 8.1)

To estimate the influence of the tested functions on the driver, different areas could also be addressed by specific questionnaires and instruments. Examples were questionnaires addressing driver behaviour (e.g. Driver Behaviour Questionnaire, DBQ), sensation seeking (e.g. Sensation Seeking Scale, SSS), and mental workload (e.g. NASA TLX). It

was suggested that part of the DBS and SSS instruments were used in order to capture the personality etc. of the participants in the background questionnaire.

The questionnaires used in TeleFOT were web-based so that the data could be stored and processed in an automated way. In order to accomplish this, an open sources tool, LimeSurvey, was chosen. A set of common questionnaires was developed and made available to the test sites in English (and was translated to the respective local languages). However, additional questions were added and additional questionnaires were developed by the individual FOTs. (See D3.2.2 Test Tools).

All questionnaires followed certain rules to assure the quality of the questionnaire. Overall though, the amount and the complexity of the questions should e.g. be limited in order to increase the probability of test participants' participation.

8.3.1.2 Interviews

An interview is a conversation between two or more people (the interviewer and the interviewee) where questions are asked by the interviewer to obtain information from the interviewee. A personal interview may be carried out in a very structured manner (involving predefined questions, predefined arrangement of questions, etc.) or in a semi-structured (e.g. carried out with the support of an interview guide) or even an unstructured manner. The particular strengths of personal interviews are that data collection can be controlled, loss of data reduced, and follow-up questions and probing procedures can be applied. This means that the personal interview allows for more in-depth information to be elicited and, hence, more in-depth understanding than do questionnaires. The main drawback of personal interviews is the reliance of (in most cases) a small and non-random sample which means that the results cannot be considered representative for the larger population in terms of statistical generalizability. Instead of the concept of 'data saturation' must be applied. 'Data saturation' often occurs after approx. 20 interviews.

In TeleFOT, structured personal (or telephone) interviews could be used as an alternative to web-based questionnaires, the reason being e.g. that not all participants may have access to the Internet. In these cases, the web-version of the questionnaire will have to

be filled in by the researchers at the organisation responsible for the local test in order to assure that the data will be available for the analysis of the results.

In addition, semi-structured individual interviews can be carried out as a complementary data collection tool in D-FOTs as well as in L-FOTs to further penetrate relevant issues and to develop, e.g. on the reasons why the test participants perceive, react to, and/or respond to the functions in a certain way. In L-FOTs the personal interviews could carried out with a sub-sample of the total sample of test participants in order to reduce the effort required.

8.3.1.3 Focus group interviews

Focus group interviews are an alternative to individual interviews. A focus group interview is a particular kind of group interview supported by a moderator. The particular strengths of focus group interviews (compared to individual interviews) are that the focus group participants can reflect not only upon the questions posed but also react to and provoke one another. Through unification across the participants, common denominators become visible and through polarisation, possible segmentation and hence differences become apparent. The drawbacks are, evidently, the reliance of a small sample. A focus group interview normally involves 8-10 participants and the total number of focus group interviews will most often not exceed 5-6.

In TeleFOT, focus group interviews was carried out as a complementary data collection tool in D-FOTs in order to further penetrate relevant issues and to develop, e.g. on the reasons why the test participants perceive, react to, and/or respond to the functions in a certain way.

8.3.2 Collecting objective data

Data collection methods to be used for the collection of objective data include measurements by logging but also activity-based diaries (or travel diaries) and observations. Again, a difference has to be made between the L-FOTs and the D-FOTs due to the limited access to data in the vehicles involved in the L-FOTs in comparison to the instrumented vehicles of the D-FOTs. However, the D-FOT data is seen as complementary to help explain and answer the research questions raised.

8.3.2.1 Logging

For the collection of objective data obtained by different sensors, different types of devices could be used. When collecting information on e.g. travelled distance, speed, etc., data can be collected by means of internal vehicle bus data (e.g. CAN), by a separate logging device, like CANcase/-alyzer, inside the specially equipped vehicles. This kind of data logging will primarily be accomplished in the D-FOTs with less than 20 vehicles, and these D-FOTs will therefore deliver more and more accurate data.

In those FOTs, were there will be no access to a vehicle bus (as when the regular systems of the driver or the automotive company are used), other loggers (Metasystems or Broadbit) have to be used and no access to the vehicle bus is available. These loggers are connected to sensors, nomadic devices or aftermarket devices, e.g. via Bluetooth. Inside the logger a GPS receiver, for the estimation of GPS Position and Time, is contained and can be equipped with other additional sensors, like accelerometers. Other more complex data acquisition systems, like driver monitoring systems, will not be installed in these cars due to the costs, lack of driver acceptance and limited data storage. Nevertheless, all loggers have to be matched to the predefined Performance Indicators, so that the needed data is available for the following evaluation subsequent to the tests.

In modern cars, all systems of the car communicate through a common bus. When, for instance, the driver turns the indicator on, the indicator switch gives a signal through the bus that the indicator light should start to flash rather than creating a closed electrical circuit. This means that the CAN-bus in a vehicle has rich information. For instance the CAN bus can provide information on speed, acceleration in three axes, fuel consumption, all driver initiated settings, which gear is engaged, steering wheel position, temperature, the presence of rain, etc. However, most of these signals is proprietary and cannot be accessed by a third party. The TeleFOT project and the L-FOTs will therefore have to rely on other means for gathering these types of objective data.

Some of the nomadic and aftermarket devices tested in the different FOTs have the capability to log data about the trips undertaken. In these cases the data will be based

on GPS. GPS provides information on the position of the device in three dimensions. From the position can be calculated speed and heading (directly in the device). If the sampling rate of the GPS data is high enough acceleration can be calculated from position data. However, depending on the research questions and the hypotheses to be addressed different sampling rates will be needed. In TeleFOT, two partners (Metasystems and Broadbit) provide loggers which use GPS and accelerometers as means for keeping track of the vehicles.

The Broadbit and Metasystems loggers are recommended if the hypotheses rely on e.g. acceleration data. Logging data directly in the nomadic and aftermarket devices, on the other hand, have the distinct advantage that no extra equipment has to be installed in the test participants' vehicles. This is both beneficial from a recruitment point of view, and it opens up the possibility for collecting objective data outside of the participants' cars. Since TeleFOT concerns all modes of transport – and not only cars – this possibility is a major advantage. Logging in the actual device also opens up for logging what functions are used, which buttons are pressed, etc. Because of access limitations to the different nomadic devices, the logging inside the device is so far only available for the Bloom device (see D3.2.2 Test Tools). With map matching it is possible to connect the GPS positions logged inside the test vehicles with maps of the road network, and thereby determining what roads the test participants have been using.

In TeleFOT FOTs data will be collected primarily by means of GPS. Direct and indirect measures will be used. In L-FOTs direct measures include e.g. measures of speed, travel time, and distance travelled (by car) whereas indirect measures include, e.g. fuel consumption. In D-FOTs the same direct measures will be used. The results from these direct measures in the D-FOTs can then be used to calibrate the algorithms for the indirect measures in the L- FOTs.

During the test runs of the D-FOTs the data to be collected is primarily objective data, obtained through different sensors mounted to the vehicle. In most cases information available on the CAN is not available (or not sufficient) for the evaluation of the chosen hypotheses, and additional data can be acquired by adding suitable sensors to the vehicle setup. However, to complete the data acquisition during the tests the objective data can

and should be supplemented by subjective data in form of questionnaires regarding e.g. usability issues and/or acceptance of functions. In L-FOTs not all relevant information can be obtained by objective data. Therefore access to this relevant information has to be ensured by additional subjective data sources, in terms of travel diaries and questionnaires.

8.3.2.2 Observations

Observation as a research method involves the act of noting and recording something, such as a phenomenon. The particular strengths of observations are that they capture on-going processes, e.g. actual behaviours of individuals, habits and routines that are not consciously reflected upon and therefore difficult to elicit by different question based data collection methods. A weakness is that the observation (if open) may have an effect on the individual's behaviour. Video-recording for later analysis may have the same negative effect, even though it can be anticipated that users, over time, will forget the camera and act according to their original habits and routines.

In TeleFOT, in-vehicle video cameras will be used in a limited number of D-FOTs. By means of these devices, video recordings data can be collected on e.g. road/traffic situation, driver behaviour, facial expressions, etc.

With access to this kind of data an evaluation of the driver's distraction and mental work load can be achieved. Therefore the eyes of the driver are to be tracked inside the view of the Cockpit-Activity-Assessment module (CAA) (see also D3.2.2.Test Tools). The algorithm of the CAA determines time of "eyes off road" and the direction of gaze. With access to these values a statement regarding 'driver distraction' can be made.

8.3.3 Additional methods and tools

8.3.3.1 Travel diaries

Collection of data by activity-based diaries involves the recording of a detailed log of how people allocate their time during the day, often focusing on particular activities, e.g. travelling. Activity diaries focusing on travel are described as 'travel diaries'. Typically these are designed to capture information on the origin and destination of a journey, departure and arrival times, mode(s) of transport used, and whether the traveller was

accompanied or not.

In TeleFOT, travel diaries will be used primarily in L-FOTs to collect information on participants' travels as well as their choice of travel modes. In addition, the travel diaries are to record the participants' use of the different functions and devices, both before, during, and after the trip. Hence, the travel diaries play a particular important role in addressing the impact areas 'efficiency' and 'mobility'. The diaries will be filled out for every trip for a defined time span, one week (see Section 8.1).

8.3.3.2 Usability evaluations

Usability is, according to ISO 9241-11, *the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use*. Usability is a fundamental quality in order to ensure use as well as user acceptance, and hence user uptake.

As part of some of the D-FOTs, but especially as a separate endeavour (cp. the TeleFOT deliverables D4.8.3 and D4.8.4), usability evaluations will be carried out in order to assess the usability of the devices and functions being tested. Usability evaluations can be achieved by, so called inspection methods and/or by empirical usability tests (Jordan 1998; Nielsen 1993).

Usability inspection is the generic name for a set of methods that are all based on evaluators inspecting a user interface. Inspection methods include e.g. heuristic evaluations and cognitive walkthrough. During a heuristic evaluation, one goes through the interface to be evaluated several times, inspects the dialogue elements (e.g. layout, menus, typography, system response times, etc.) and compares them with a list of generally acknowledged usability principles – heuristics – as well as specific category ones. Other principles may evidently also be considered. The output from a heuristic evaluation is a list of usability problems in the interface with references to those usability principles that were considered violated. If heuristic evaluations are to be carried out, it is advisable that the evaluation is carried out by more than one evaluator, preferably by more than five (Nielsen & Mack, 1994).

Cognitive walkthrough involves one or a group of evaluators inspecting a user interface by going through a set of tasks and evaluate its coherence and ease of learning. The user interface is often presented in the form of a paper mock-up, a working prototype, or a fully developed interface. The input to the walkthrough also include the user profile, especially the users' knowledge of the task domain and of the interface, and the tasks to be evaluated, the correct (intended) way of accomplishing the tasks, and the way the system/interface responds to different input. The analysis phase consists of examining each action in the solution path by asking following four questions: Will the users try to achieve the right effect? Will the user notice that the correct action is available? Will the user associate the correct action with the effect to be achieved? and
If the correct action is performed, will the user see that progress is being made toward solution of the task? The outcome of the analysis is a description of successes (when the user is assumed to be able to succeed) and failures (when the user is assumed not to succeed) and the possible reasons why (Nielsen & Mack, 1994).

Usability tests normally involves a sample of individuals (normally 12-20) representative for the user population who are asked to complete certain tasks (in contrast to the methods mentioned above where experts use different methods to evaluate a user interface according to certain criteria). During the tests, the test participant is asked to ‘think aloud’ in order collect data on, e.g. problem solving strategies. Data is also collected by observations of the test participants’ behaviour and interaction with the product/system. Based on the usability definition, evaluation criteria include effectiveness, efficiency, and satisfaction and relevant derived measures, e.g. percentage of tasks completed, task completion times, number of errors, error recovery times, and ratings of ease-of-use. These measures can be used in order to compare different solutions (benchmarking) as well as assess whether or not a design solution meets (or does not meet) specified requirements (Jordan 1998; Nielsen 1994).

In TeleFOT, the usability tests should be designed and accomplished according to the standard procedures for evaluation by theoretical or empirical methods.

8.4 Data storage and management

The objective data (logging, video-recordings, as well as travel diaries) and the subjective data (questionnaires) from both L-FOTs and D-FOTs were stored on a central database server (see D3.2.2 Test tools) where they were accessible for the evaluation team. If interviews and/or focus group interviews are carried out, these data will be treated separately.

In the L-FOTs, objective data was collected via special data loggers. These loggers gather all available data, e.g. acceleration, vehicle speed and position, and send the information via GPRS or GSM connection either directly to the central TeleFOT data base, (if the format is correct) or to the database of the provider of the logging device. In this case the data is converted into the central database format and afterwards forwarded to the central TeleFOT database.

In the D-FOTs the amount of data to be logged is larger than is the case in the L-FOTs (due to video data and the logging of other parameters distributed by the CAN of the equipped vehicles). Therefore this data is stored at the company/organisation that is performing the tests. Data to be included in the central database is extracted from the log-files and converted into the central database format. Following to the conversion, the data is forwarded to the central database via a web connection.

The data flow in relation to the Data Base (DB) can be described as in Figure 8.2, where the different processing steps at the central storage are highlighted.

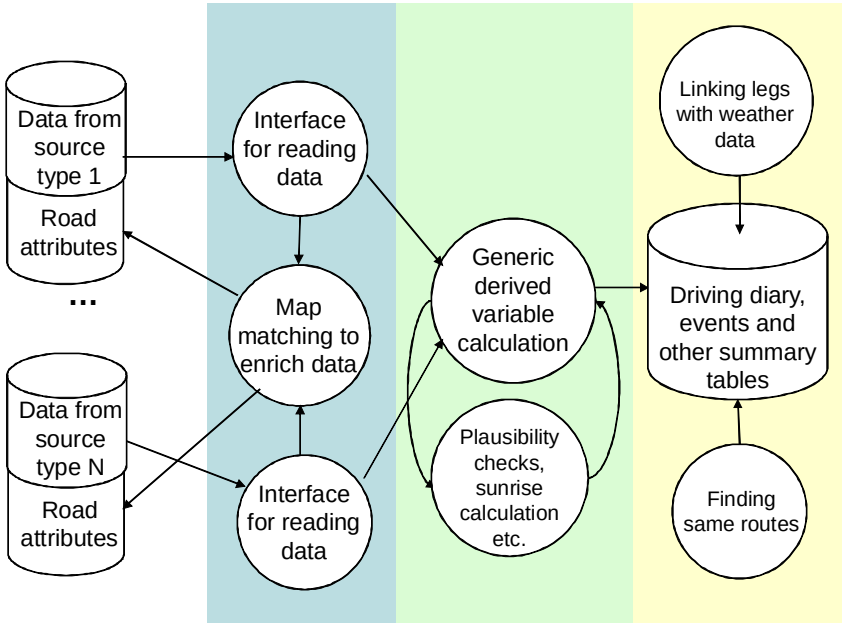

Figure 8.2 Data processing steps at the central storage

The data was pre-processed before stored on the central database server, so that a unique basis for the following evaluation could be guaranteed. With the pre-processing different parameters to be evaluated (e.g. trip length, trip duration etc.) could be extracted from the log files. The pre-processing have to be carried out according to predefined reference values, e.g. 'engine start', 'engine stop', so that a reliable evaluation of the data can be carried out.

Different types of summary table were compiled and examples are:

- Driving diaries, where each leg is described by a long list (>200) of derived variables and indicators such as
  - the number of hard brakings
  - the total distance driven on a road type x
  - weather conditions,
  - etc.
- Event lists, e.g. detected speeding events
- List of most common legs

Table 8.1 An example of a summary table report

Speed limit	Driving according to the speed limit [%]	1–9 km/h speeding [%]	10–19 km/h speeding [%]	>20 km/h speeding [%]	Total driving [km] = 100%
30–40	84.0	13.8	2.1	0.0	420
50–70	79.5	17.0	3.4	0.1	1241
80–90	49.6	31.2	14.4	4.8	515
100–120	69.9	23.7	5.7	0.7	1119
Speed limit uncertain (Finnish Digiroad data example)					644
Total					3938

The reasons for these database arrangements were to:

- Make access to data easier
- Make possible collaborative analyses of all impact areas of all FOTs and not just a single FOT
- Create a final storage and make a record of data ownership, study designs, user information tables and all relevant information for analysis
- Convert data into a common format (where reasonable)
- Harmonize post-processing to produce summary tables consisting of pre-defined indicators.
- Let analysts re-create their analysis databases using database dump files to be downloaded from the central storage.

The Figure 8.3 shows a general overview of the data flow.

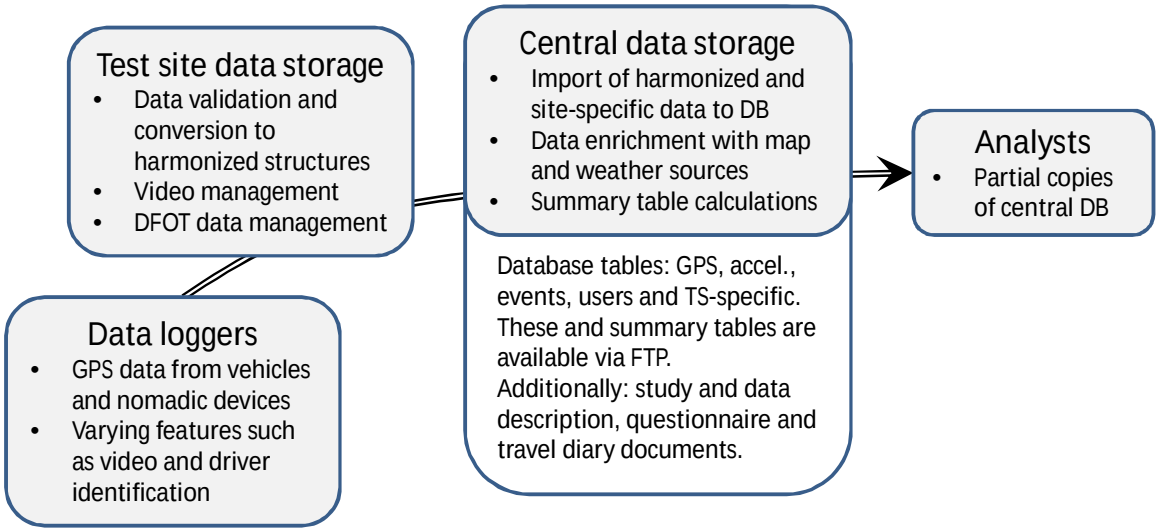

Figure 8.3 General Data Flow Overview

Some lessons have been learned along the life time of the project:

- FOT logging requires much resources, consideration and testing. Long-time driving statistics can be studied with common in-vehicle loggers but detailed event analyses benefit from video and radar
- Efficient data validation is important to avoid having to repeat tests
- Compiling metadata (documentation and supporting tables) is absolutely necessary if FOT data is to be easily shared with others. Also test sites need to document topics when still remembered.
- Harmonizing logging and post-processing enables analysts to more easily cover several test sites.
- Common processing reduces individual work and it also ensures comparability.
- Summarised/aggregated content gives an index into raw data. General indicators can be calculated from all FOTs

9 PREPARING FOR THE TEST

9.1 Pilot tests

Conducting a pilot study is necessary to prepare for the deployment of the different FOTs, whether L-FOTs or D-FOTs, and to support the design of the relevant tools for the evaluation process (See FESTA Handbook; Saad, 1997; Saad and Dionisio, 2007). It represents a relevant step for the mobilisation and dialogue between the various teams involved in the FOTs, hereby promoting a common framework and consensus for the evaluation process.

Pilot tests in FOTs should address the following levels of analysis:

- On the first level the pilot tests have to check the technical function of the data collection systems in actual driving situations. They should enable to identify potential problems of sensor calibration and thus to establish the periodicity of maintenance procedures during the FOT, and to validate the data collection procedure (from data acquisition, data transmission to data storage). The technical teams involved in the FOT should be in charge of these pilot field tests. The first pilot test should be executed with the full set of objective data. The pilot test must proof which of the full data set and full test procedures that can be adapted to the FOT and which can not.
- The second level consists of testing the feasibility of the overall evaluation process – from the selection of participants through to data collection. It is, thus, a final rehearsal before the deployment of the actual FOT. It enables a check of the communication process between the various teams involved in the practical deployment of the FOT and of the robustness of the technical tools designed for data collection and transmission.

9.2 Informing the participants

All participants, in L-FOTs and in D-FOTs, need to be informed in advance about the purpose of the specific FOT, of any risks they may incur, the costs that are covered and

not covered (L-FOTs only), whom to contact in case of breakdown (L-FOTs only), etc. For this purpose and for legal reasons, there is a need to formalise the arrangement between the involved organisations and participants themselves.

In order to obtain a valid consent from the participant (as well as their employer if the cars used are company cars) in the FOT to log the data and allow for safe participation in the FOT, information on the test procedure and setup must be comprehensive. For example, the information should include which kind of data is being logged, who will have access to it, how to deal with malfunctions, responsibilities, etc. This is as important in the L-FOT as in the D-FOT.

This information can be provided in the form of a written manual, as well as through personal briefings or presentations. The possibility to ask questions at any time later during the FOT should be provided.

10 ETHICAL AND LEGAL ISSUES

All studies involving human participants must take into account legal and ethical aspects. FOTs in which vehicles are used may jeopardise people's safety, security as well as privacy. As a consequence the organisations involved in an FOT may be held responsible of any harm incurred by persons, whether actively involved in the FOT or not. Thus ethical and legal expertise must be involved at an early planning stage, well before the execution of the specific FOT and until the final analysis of the data. As a general rule it is advised to perform a thorough risk assessment and risk management along the life of the FOT. (Furthermore, if an FOT should involve different countries, legal experts should evaluate the different regulations or interpretations in the countries involved.)

In the following a minimum set of actions to be performed before carrying out the FOTs in TeleFOT is described. In addition the organisation responsible for the FOT should evaluate the need to complete a risk assessment and a system safety assessment.

10.1 Administrative matters

All participants of an FOT should be selected and enrolled in a proper process. This should include at least the following steps:

- Select only drivers with valid driving license and active insurance coverage, especially when driving their own vehicles. If necessary, a specific additional insurance must be drawn up.
- Inform drivers about the purpose of the test and the risks incurred (see Chapter 8).
- Specify, enforce and eventually endorse in a specific contract, the following points:
    - payments to the driver (monetary or not);
    - responsibilities;
    - costs covered; and
    - coverage of damages to own vehicle and occupants.

10.2 Ethical issues

When planning the FOT, it is vital to check whether or not a review by an ethical board is necessary or not[5] and to plan for the time needed for formulating an application and its approval. Also, different organisations, e.g. different universities, may have internal ethical committees that must review experiments involving human beings.

Each FOT is advised to create a specific ethical committee involving technicians, researchers, and legal and ethical experts before planning and running the FOT. This committee should ensure that all necessary permissions for running the FOT have been acquired.

10.3 Privacy

All FOT test sites must make certain that the privacy of the participants is secured. The organisation responsible for the local FOT should address data protection in the way required by the laws of each European member state. Data retrieved during an FOT may enable the identification of the persons involved, in particular in case that video is recorded during the test, and furthermore when passengers are also covered. Thus, a proper process of data anonymisation should be defined: masking of the IDs and/or disabling visual identification of the persons. Also, if data are logged by a data logger a technical option (a “red button”) may be needed that allows the driver to delete all data in case of an accident. Otherwise the legal authorities could require and make use the logged data to sue the driver. If the driver has no access to such a deleting function, a legal clarification of this case is recommended. Regarding video acquisition, added

[5] In Sweden, a new law has been in force since 1 January 2004 which deals with vetting the ethics of research that involves humans. It encompasses research involving living persons, but it also covers such areas as, e.g. research on the deceased, biological material from people. The most important change since the beginning of 2004 is the establishment as independent authorities of one central ethical vetting board and six local ones. A fee is charged for each application. The fee is to be paid by the person principally responsible for the research. For more information, see http://www.epn.se

difficulties may arise from the recordings are made in private areas or nearby sensitive areas, such as borders, government buildings, prisons, and schools. This should be taken into consideration in the planning of the FOT, or in the anonymisation process. Any personal information should be maintained in a separate and protected database.

10.4 Security

All test sites must make certain that the security of the participants is apprehended. The organisers of the FOT is recommended to follow a structured risk management procedure in terms of planning, assessing, and managing in case the organisation responsible for the FOT will need to be able to demonstrate that the procedure has been implemented and followed. Different organisations will normally have a safety management process for this.

In the case that new equipment taken on-board the vehicles involved in the FOT, their proper installation should be verified (eventually also in the framework of a risk management procedure). In case that major modifications are needed, the approval for on road use should be obtained by the delegated national authority. In particular the FOT should verify that modifications do not yield:

- Reduced visibility;
- Modification to existing equipment functioning;
- Added radio or electric interferences;
- Added distractions to the driver;
- Reduced protection of vehicle occupants in case of crash.

11 IMPLEMENTATION OF STRATEGY

11.1 Appointment of an Evaluation Manager

The strategy for assessment and evaluation will be implemented step by step and FOT by FOT as more and more details of the actual functions/systems as well as the local test site characteristics are made known. In this process a necessary prerequisite is the existence of a test site *Evaluation Manager* with ample knowledge and experience in the area of how an assessment and evaluation process would be initiated, designed and conducted. This person could be the *Test Site Manager* if the competence profile is the right one, otherwise a dedicated Evaluation Manager must be appointed.

The main task of the Evaluation Manager is to make certain that the FOTs to be performed at the local TeleFOT Test Sites will follow the principles of the TeleFOT assessment and evaluation strategy as set out in the deliverable D2.2.2... The most important argument for this step is that only in this way a European approach in the evaluation work can be guaranteed. Every impact area must be addressed in a way which will makes possible a comparison between the impacts of function/systems when implemented in different contexts, i.e. in different FOTs and countries.

11.2 Collaboration is necessary

A close collaboration (from time to time quite intensive) between the Evaluation Managers and the team responsible for the strategy is foreseen. In the context of TeleFOT it goes hand in hand with the physical arrangements of the FOTs (SP3) and, where relevant, follows the FESTA FOT Implementation Plan. However, the most central actions of the Evaluation Manager is to make sure that the study design and the evaluation plans of each FOT (dedicated to a specific or a combination of functions/systems) will be developed in a scientifically sound and acceptable way (related to SP2). Furthermore the preparation of the evaluation work and the support needed when the analysis work starts imply strong collaboration between SP2 and SP4.

12 CONCLUSIONS

This deliverable, D2.2.2. *Testing and Evaluation Strategy II,* provides a verified and validated description of the recommended *general strategy* to be applied in and across the different FOTs, both Large Scale and Detailed FOTs. The general strategy includes the overall, recommended approaches for generating research questions and hypotheses, for choosing study design, for recruiting and choosing participants, as well as a description of what type of data is to be collected and recommendations for data collection methods and procedures to be used.

The TeleFOT project is characterised by a mix of, so called, Large Scale and Detailed FOTs. The L-FOT were run as naturalistic driving tests (to as high an extent as possible), with a large number of vehicles and participants taking part in the test over a longer period of time, while the D-FOTs were seen as complementary to the LFOTs and as 'controlled experiment'. The differences have implications for study design, for the data that can be collected, and by what methods.

Consistent across L-FOTs and D-FOTs, however, a study design involving baseline/control conditions compared with experimental conditions was applied, both subjective and objective data was collected in order to address a set of common research questions and hypotheses, and data was collected pre-test, during, and post-test. Legal and ethical matters, including privacy and security, were addressed in the same rigorous way across all test sites.

Furthermore several of the test sites evaluated combinations of functions. Even though this mirrors actual use situations it adds complexity to the project, the formulation of research questions and hypotheses, the study design, the data collection procedures, as well as the analysis of the results.

For generating the relevant research questions and hypotheses, an integrated top-down and bottom-up approach was developed within the project. The top-down approach is

driven by the issues of relevance to the impact areas irrespective of the system functionality while the bottom-up approach is driven by the different functions to be tested. This will allow for the necessary flexibility but also for consistency when new functions are added for evaluation in the later stages of the TeleFOT project. Furthermore, the approach but specifically the top-down hypotheses provide an important input to FOT evaluations in general (also outside the transportation sector).

REFERENCES AND OTHER LITERATURE

Cohen, J. (1992). A power primer. Psychological Bulletin, 112, pp. 155-159.

Reports from the FESTA project, all of which are available at http://www.its.leeds.ac.uk/festa/downloads.php

FOT-NeT web page (www.for-net.eu)

International Organization for Standardization (1998): ISO 9241-11:1998. Ergonomic requirements for office work with visual display terminals (VDTs). Part 11: Guidance on usability.

Jordan, P. (1998): An introduction to usability. Taylor & Francis Ltd, London.

Nielsen, J. (1993): Usability engineering. Academic Press, CA.

Nielsen, J. & Mack, R. (eds.) (1994): Usability inspection methods. John Wiley & Sons, N.Y.

Michon, J. A. (1985). A critical view of driver behaviour models. What do we know, what should we do? In L. Evans & R. Schwing (Eds.), Human behaviour and traffic safety (pp. 485-525). Plenum press, N.Y.

Popkin, S., Wilson, B. & Howarth, H. (2003): Drowsy driver warning system field operational test: experimental design considerations. Volpe National Transportation Systems Center, Cambridge, MA.

Regan M. A., Young, K.L., Triggs, T.J., Tomasevic, N., Mitsopoulos, E., Tierney, P., Healy, D., Tingvall, C. & Stephan, K. (2006): Impact on driving performance of intelligent speed adaptation, following distance warning and seatbelt reminder systems: Key findings from the TAC SafeCar project. IEE Intell. Transp. Systems, Volume 153, No. 1, March 2006.

Rumar, K. (1993): Road user needs,. In Parkes, A. S. Franzen (Eds): Driving Future Vehicles, Taylor &
Francis, New York, pp. 41-48.

van der Laan, J.D., Heino, A., & De Waard, D. (1997). A simple procedure for the assessment of acceptance of advanced transport telematics. Transportation Research - Part C: Emerging Technologies, 5, pp. 1-10