Decision-making based on measurements of behavioural safety
An explorative study into using Situation Awareness measurements for a specific problem as input to system-level decisions

R.J. van der Meij
Decision-making based on measurements of behavioural safety

An explorative study into using Situation Awareness measurements for a specific problem as input to system-level decisions.

By

R.J. van der Meij

in partial fulfilment of the requirements for the degree of

Master of Science
in Engineering And Policy analysis

at the Delft University of Technology,
to be defended publicly on November 25, 2013

Student number: 1227874
Thesis committee: Prof. mr. dr. J.A. de Bruijn, TU Delft
Dr. ir. S.A. Meijer, TU Delft
Dr. S. Sillem, TU Delft
Ir. J.J.W. van Luipen MBR, ProRail

An electronic version of this thesis is available at http://repository.tudelft.nl/.
This work in this thesis was supported by ProRail. Their cooperation is hereby gratefully acknowledged.

Cover picture: Eye tracking measurements in the full scope simulator of the NSO.

Photographer: Hermine van Broggel
Preface

This thesis concludes my master of Engineering and Policy Analysis at Delft University of Technology. Writing this thesis would not have been possible without a couple of persons. I would like to take this opportunity to thank them all.

First of all, I would like to thank Sebastiaan Meijer, who was my first supervisor and helped me structure my project and pushed me into the right direction when I was lost. Thanks to him I got introduced to Jelle van Luipen, who was my daily supervisor at ProRail. I remember the first meeting with him on a Monday morning; a half hour before our meeting he did not know about my existence. Unprepared we had a meeting after which he gave me the trust of writing my thesis at ProRail and be part of an exceptional project. I am really thankful for this trust he gave me. I would also like to thank Simone Sillem for reading my thesis thoroughly and giving very detailed feedback on my writing. This helped me a lot to make my thesis scientifically sound. Finally, I would like to thank Hans de Bruijn, who helped me understand the content by asking lots of questions.

Furthermore, I would like to thank Julia Lo, who helped me with the initial steps of my thesis. She introduced me into SA and gave me lots of feedback on my SA design. Also, I would like to thank Rolf Zon of the NLR. He helped me with designing the SA measurements by giving me useful feedback and was always available for consultation. Furthermore I would like to thank Jop Groeneweg, for the interesting discussion and giving me feedback and input to think about. I would also like to thank the colleagues at ProRail for their support and ‘gezelligheid’. A special thank to Marjan Schotsman, who read and corrected my thesis.

There are some people who I would like to thank, who were not directly involved with my thesis, but without them I could not have finished it.

First of all, my brother and sister, who supported me all the years of my study. I would also like to thank my friends in Delft. Thanks to you I have had a great time in Delft.

A special thank to my parents, who supported me through the sometimes-difficult years of my study. They were always available for consultation and support when I did not pass an exam, and the celebrations when I passed exams. Without their support my studies would not have been possible.

Finally, I would like to thank my girlfriend. She had to listen to all my frustrations during my study time. She always managed to create a smile on my face and took great care of me when I was stressed out. Without her, I would not have been this resilient.

René van der Meij

Delft, November 2013
Abstract

ProRail is trying to increase the capacity of the rail infrastructure since the need for capacity is still growing. One of the ways to do this is by implementing DDDR (Dynamic double non-interrupted braking), which is part of the DSSU project. By implementing DDDR the expectations of the status of the signs of the train drivers might change. Therefore ProRail wants to investigate if implementing DDDR will have a negatively impact on safety. In this thesis is investigated how situation awareness (SA) can be measured and if it is a useful safety indicator to take safety decisions.

SA measurements can be used for testing if the perception, comprehension, and expectations of the signs in the environment are influenced by DDDR. Decisions are made based on the SA; an informed, conscious decision cannot be made without a correct SA. SA is a requirement, not a guarantee, for making informed, conscious decisions.

Three factors are identified that influence the acceptability of the SA measurements. First, the safety culture has an impact on the acceptability. It is more likely that negative results will be neglected when companies have a bad safety culture. Second, strategic decision-making has an impact on the acceptability. In all phases of decision making; goal formulation, alternative generation, and evaluation and selection, should safety be incorporated. Other interest, such as economic interest, might get the upper hand when safety is not incorporated in all three steps. Third, the use of the Common Safety Method (CSM) is a formal requirement for the acceptability of the results. The CSM has three requirements for measuring safety: the safety of a new design should be comparable with the current situation, since the safety should be at least as good as the current situation, the method should correctly reflect the system, the used parameters and results should be accurate enough to serve as a robust decision-support.

The best method for comparing designs is a controlled before and after study. For the best comparison between designs, the researcher can use a simulator in order that he can exactly determine which elements are changed in the situation. Since people can show different behaviour when they are being tested, they should not be informed when they are being tested. Arbitrary measurement moments can be chosen to enhance that participants do not show different behaviour when they are studied; the participant cannot prepare for the measurements. To make the outcome of the study more robust, a mixed method design can be used; the conclusion of a study is more robust if all the methods have the same outcome.

Situation awareness global assessment technique (SAGAT) is the best-known direct and objective measurement method for SA; this is the recommended method for determining SA for train drivers. By combining this method with the Mission awareness rating scale (MARS), a more robust answer can be found. In the DSSU study, SAGAT was capable to identify differences in SA between scenarios. MARS was not capable to show significant differences between scenarios, because of the low amount of participants. The SA study was combined with a workload study via eye-tracking and biometric measurements, and a qualitative time-to Signal Passed at Danger (TT-STS) study. These studies combined concluded that the probability of a Signal Passed at Danger (SPAD) decreases in DSSU compared to the current situation.
Generalizability should be taken into account before a study is executed if in the future new SA measurements will be held. In the DSSU project the outcome of the study only applied to DSSU. By using the identified elements in chapter 7, scenarios can be made which are expected to have a bad influence on the SA. The worst-case scenario for SA can be used for measuring SA in the project. If this scenario is not applicable for the researched project, management has to decide whether they want results that are only applicable for the project or generic results, which are applicable on a system-level.

The SA measurements showed to be a useful indicator for safety when it is combined with TT-STS measurements. The combination of the measurements gave a detailed insight in changed behaviour and awareness in the new situation. SA cannot be used as a single safety indicator, since SA is a requirement, and not a guarantee, for making informed, conscious, and safe decisions.
Contents

Preface .................................................................................................................................................. i

Abstract ............................................................................................................................................... iii

Contents ............................................................................................................................................... v

1 Introduction .................................................................................................................................. 1
  1.1 Background ............................................................................................................................... 1
  1.2 Problem statement ..................................................................................................................... 4
  1.3 Research goals ........................................................................................................................... 4
  1.4 Research questions .................................................................................................................... 5
  1.5 Research Methods and Outline of the report ........................................................................... 8
  1.6 Scope ........................................................................................................................................ 10

2 Situation Awareness ..................................................................................................................... 13
  2.1 Definition .................................................................................................................................. 13
    2.1.1 Level 1: Perception ............................................................................................................... 14
    2.1.2 Level 2: Comprehension ...................................................................................................... 14
    2.1.3 Level 3: Projection ................................................................................................................. 15
  2.2 Mental model ............................................................................................................................. 15
  2.3 Goals ......................................................................................................................................... 15
  2.4 Expectations and Automaticity ................................................................................................... 16
  2.5 Safety ....................................................................................................................................... 16
  2.6 Conclusion ................................................................................................................................ 19

3 Decision-making ............................................................................................................................. 21
  3.1 Safety culture ............................................................................................................................. 21
  3.2 Strategic decision-making ........................................................................................................... 24
  3.3 Common safety method ............................................................................................................. 27
    3.3.1 System definition .................................................................................................................. 27
    3.3.2 Hazard identification .......................................................................................................... 27
    3.3.3 Risk acceptance strategy ..................................................................................................... 28
    3.3.4 Risk evaluation .................................................................................................................... 29
    3.3.5 Safety requirements ............................................................................................................. 29
    3.3.6 Demonstration of meeting safety requirements ................................................................... 29
  3.4 Conclusion ................................................................................................................................ 29

4 Study design ................................................................................................................................... 33
  4.1 General study designs ................................................................................................................. 33
  4.2 Human factor study .................................................................................................................... 36
  4.3 Measurement points .................................................................................................................... 36
  4.4 Mixed method design .................................................................................................................. 36
  4.5 Conclusion .................................................................................................................................. 37

5 SA measurement methods .............................................................................................................. 39
5.1 Overview ......................................................................................................................... 39
  5.1.1 SA requirement analysis ......................................................................................... 40
  5.1.2 Freeze probe techniques ....................................................................................... 40
  5.1.3 Real-time probe techniques ................................................................................... 40
  5.1.4 Self-rating techniques ......................................................................................... 41
  5.1.5 Observer-rating techniques ................................................................................... 41
  5.1.6 Performance measures ......................................................................................... 41
  5.1.7 Process indices ..................................................................................................... 41

5.2 GDTA ............................................................................................................................. 41

5.3 SAGAT ............................................................................................................................. 42

5.4 SPAM ............................................................................................................................... 43

5.5 MARS ............................................................................................................................... 44

5.6 SABARS .......................................................................................................................... 45

5.7 TT-STS ............................................................................................................................. 46

5.8 Eye track ............................................................................................................................ 47

5.9 Measurement environment ............................................................................................. 48

5.10 Conclusion ...................................................................................................................... 48

6 Case study: DSSU ................................................................................................................. 49
  6.1 DSSU description ...................................................................................................... 49
  6.2 Simulator ..................................................................................................................... 51
  6.3 Scenarios ..................................................................................................................... 52
  6.4 Participants .................................................................................................................. 53
  6.5 SA measurement set up ............................................................................................. 53
  6.6 Data processing .......................................................................................................... 54
  6.7 Data analysis ............................................................................................................... 54
  6.8 MARS ............................................................................................................................. 57
  6.9 SAGAT vs. MARS ........................................................................................................ 62
  6.10 Comments .................................................................................................................. 62
  6.11 MATRICS .................................................................................................................... 63
  6.12 Conclusion .................................................................................................................. 64

7 Generalizability and Transferability .................................................................................... 67
  7.1 Determination of SA in other projects ......................................................................... 67
  7.2 Generalization theory ................................................................................................. 68
  7.3 Analytical generalization ............................................................................................ 69
  7.4 Transferability .............................................................................................................. 70
  7.5 Relation between generalizability and precision of outcomes ..................................... 70
  7.6 Usability of the methods ............................................................................................. 71
  7.7 Improving generalization ............................................................................................ 72
  7.8 Elements for generalization ....................................................................................... 73
  7.9 Generalizability of the results ..................................................................................... 75
  7.10 Reflection .................................................................................................................... 78
  7.11 Consideration of strategic management .................................................................... 79
  7.12 Comments .................................................................................................................. 80
  7.13 Conclusion .................................................................................................................. 80
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>Conclusion</td>
<td>83</td>
</tr>
<tr>
<td>8.1</td>
<td>Conclusion</td>
<td>83</td>
</tr>
<tr>
<td>8.2</td>
<td>Recommendations</td>
<td>88</td>
</tr>
<tr>
<td>9</td>
<td>Reflection</td>
<td>89</td>
</tr>
<tr>
<td>9.1</td>
<td>How did my research contribute to current research?</td>
<td>89</td>
</tr>
<tr>
<td>9.2</td>
<td>How useful is measuring SA of train driver?</td>
<td>89</td>
</tr>
<tr>
<td>9.3</td>
<td>What did I learn from my thesis?</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>Appendix</td>
<td>91</td>
</tr>
<tr>
<td>Appendix A: Glossary</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>Appendix B: List of figures</td>
<td>92</td>
<td></td>
</tr>
<tr>
<td>Appendix C: List of Tables</td>
<td>93</td>
<td></td>
</tr>
<tr>
<td>Appendix D: SAGAT question forms DSSU</td>
<td>95</td>
<td></td>
</tr>
<tr>
<td>Bibliography</td>
<td>103</td>
<td></td>
</tr>
</tbody>
</table>
1 Introduction

1.1 Background

The Dutch railway network is one of the busiest networks in the world (CBS, 2009, Goverde, 2005). ProRail is responsible for the construction, maintenance, management and safety of the Dutch Railway network (ProRail, 2013c). The need for capacity is still growing (OVPro.nl, 2012), also the Dutch government wants to increase passenger transportation by rail by 5% annually (Ministry of Infrastructure and the Environment, 2007). The KiM (Kennisinstituut voor Mobiliteitsbeleid) expects a growth of 35% to 70% during the rush hour of passenger transportation on the rail between 2006 and 2020 (Ministry of Infrastructure and the Environment, 2007). Because of the growth of the use of rail transport in the previous years it is expected that, by continuing the current operating principle, the system is at its limit of capacity (Ministry of Infrastructure and the Environment, 2013). Increasing capacity can be done via changes in the processes or in the infrastructure. ProRail tries to use the same infrastructure but change the way it is used (van den Hoogen and Meijer, 2012).

There are several ways to generate extra capacity without adding new rail tracks; one of the ways is to use shorter signal blocks. This is one of the measures of the DSSU-project (in Dutch: DoorStroom Station Utrecht). Among other components, two components in a timetable determine the amount of time needed between two trains, and thus determine the maximal capacity. First, the headway, which is the technical minimal time between two trains determined by the constraints of the signalling system. Second, the buffer time between two trains, which is needed to dampen differences in driving behaviour of both trains. Implementing the extra blocks can reduce the headway time, i.e. trains can drive closer to each other. When trains drive closer to each other the effect of a delayed train on the punctuality of other trains is reduced. Trains can never be in the same block to avoid collisions. When two trains are behind a delayed train, the distance between the delayed train and the second train is 2 blocks. If the lengths of blocks are reduced from 1300 meters to 400 meters, the rear train has a distance reduction of 1800 meters. Furthermore, if the blocks are smaller, the first train will make the block earlier available for the second train.

To make the shorter signal blocks possible a new braking system should be implemented called DDR (Dynamisch Dubbele Doorgaande Remming, freely translated: dynamic double non-interrupted braking). In the old signal system the length of the blocks are based on the maximal breaking distance of a train. In the current signal system a train normally gets a green light; the maximum speed is allowed. A red sign is placed if there is an obstruction; the train driver has to stop before the red light. Before the red sign there is always a yellow sign displayed; at the next sign the reduce speed till 40 km/h. and expect to stop at the next signal. Multiple yellow signs can occur in sequence if the obstruction is moving (e.g. a train is moving from one block to the other). In the DDR system the green and red light keep the same function. The difference will be in the use of the yellow signs in normal conditions. Since the blocks are shorter, the yellow sign has to be placed earlier. Extra signs are used, first one or more yellow signs with a (blinking) number, then a yellow sign as in the conventional system. A yellow sign with a blinking number stands for: Braking should not be interrupted if next sign allows a lower speed than the blinking number. In DDR the yellow + blinking number sign can occur in succession of each other.
Before this measure will be implemented, ProRail wants to be sure that it does not influence the safety negatively. Previous research suggested that DDR might have negative consequences for the behaviour, performance, alertness and fatigue of the train driver. Especially the expectations of the train drivers of the next sign can change (Zon et al., 2013). By measuring the situation awareness (SA) of train drivers, it can be tested if the perception, comprehension, and expectations of the elements in the environment are influenced by DDR. Endsley described SA as follows: “The perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future” (Endsley, 1988a, p. 97). In the research of van den Top (2010), about risk control measures in the railways, is stated that SA is a appropriate method for comparing designs for human factors: “Situation awareness should be a main assessment criterion to see how a proposed change influences the operator’s ability to understand the situation” (van den Top, 2010, p. 259).

SA can have an influence on the safety, but the actual relation has still to be determined (Salmon and Stanton, 2012). If SA is used for making safety decisions, this relation has to be determined first. Implementing new signalling configurations on the railway tracks can have an impact on the SA of train drivers. This can be a reduction of SA, because new signalling can be confusing, or because train drivers get too much information and therefore might miss important information. The contrary can also happen: an increase of SA. When the new signalling system is implemented train drivers might be extra alert and therefore have a better SA, or because of the extra information they get they might have a better SA. A reduction of SA can have a bad influence on the safety, but it does not necessarily mean that a train driver takes risks that he is unaware of. A train driver can also be extra cautious when he is unsure about his situation and therefore drive slower than needed. This is also an unwanted situation: too much breaking can give delays. But bear in mind, in this situation the train driver is aware of the fact that he is unaware of the situation! Generally speaking, a lower SA leads to a lower safety: Endsley and Kiris stated that low SA leads to slower detection of problems and more time was needed to investigate problems and conducting activities when a problem is detected (Endsley and Kiris, 1995). Smith and Hancock link SA to the capacity to generate consciousness: “We argue that SA is more than performance. More fundamentally, it is the capacity to direct consciousness to generate competent performance given a particular situation as it unfolds” (Smith and Hancock, 1995, p.138). In both statements SA is a prerequisite for safe performance. Furthermore, SA helps to form a mental picture of the situation. Van den Top stated:” without a correct mental picture of what is going on, it is not possible to take the right decisions” (van den Top, 2010, p.138). Therefore it is very interesting to know what the SA of the train driver is in various signalling configurations along the tracks and display technologies in the cabin of the driver. Since ProRail is responsible for the safety of the Dutch railway network, it is important to know how new signalling configurations or new displays containing information provided (partly) by ProRail, influences the train driver’s SA.

Measuring SA
To be able to make decisions on SA, it first has to be measured. SA measurements have already been applied in several areas such as drivers (Walker et al., 2008), pilots (Bell and Waag, 1997), and air traffic control (Endsley et al., 2000b). A lot of different methods are available for measuring SA (Salmon et al., 2006):
1) SA requirement analysis → e.g. GDTA
2) Freeze probe techniques → e.g. SAGAT, SALSA, SACRI
3) Real-time probe techniques → e.g. SPAM, SASHA
4) Self-rating techniques → e.g. SART, SARS, CARS, MARS, QUASA
5) Observer-rating techniques → e.g. SABARS
6) Performance measures → e.g. kills, hits, hazard detection
7) Process indices → e.g. eye tracker, heartbeat

From these methods all the concepts are very clear given, but it is still unclear which method is best to choose. This might be one method, or methods can be combined and adjusted. Measurements should also be “future proof”; these SA-values should still be useful when, in the future, small changes occur at the tracks, signalling or cabins.

Furthermore, the SA measurements are based on the goal-directed task analysis (GDTA). The GDTA was developed by combining Dutch and British documents of train driver’s occupational competences, which was converted into tasks. After that, a Dutch train driver and an instructor separately validated the GDTA. It should be determined if there is a need of new measurements if these goals change. Also, the measurements methods are explained in concept, but an actual SA measurement cannot be found.

In this thesis a comparison is made between the methods in which the pros and cons are described, and the explanation of the measurement methods will be given.

**Acceptability of the results**

Both project and strategic management should be able to make safety-based decisions on the SA measurement. These parties have different interest in the results. Project management decide on the safety for this particular project (DSSU); they want to know if DSSU with DDDR can be implemented safely, i.e. on a project-level. Strategic management decide on the appliance of the outcomes of the measurements in general; they want to know how these measurement results apply to other locations and other settings, i.e. on a system-level. Therefore strategic management is interested in how the results can be generalized, generalizability or transferability of the results should be taken into account in the study design.

ProRail experts, known as “railverkeerskundigen” (RVT), should approve changes before it can be built and used. They should be able to conclude something from an SA measurement. But there are some problems with measuring SA: The measurements do not give exact values of SA and the direct influence on the safety is also not clear. They want to know if measuring SA is enough to draw safety conclusions from, or if extra data is needed. Human factors are not structurally implemented in their day to day decisions. It should be investigated how they can implement SA measurements in their decisions.

Finally, ProRail uses the Common Safety Method (CSM) as their risk assessment method, a risk analysis method developed to harmonize the methods of the railway sector in Europe (European Railway Agency, 2012). The requirements of the Common Safety Method (CSM) are a prerequisite for the SA method.
In this thesis research will be investigated how SA can be measured by ProRail (together with NS Dutch Railways) and how they can use this data to make decisions on the safety of changes in this project (DSSU) and other, yet undefined, projects.

**Long-term effects**
A problem with new techniques or situations is the fact that train drivers do not have experience with these circumstances. SA measurements are often executed by comparing the current situation with a new situation (i.e. a measurement without and with a new technique implemented). Previously stored knowledge has a big influence on how people see the world, there is a disjoint between the world as they see it and the world as is (Dekker et al., 2009). The SA of a train driver in a new situation should therefore be lower than in the old situation; they expect the old situation and do not see the changes that occurred. On the other hand, they will be extra alert, if they notice that they are in a different situation. Only after a long period after the implementation will the new situation be normal. How much time this is should be determined. This again can have two outcomes; the new situation can be very normal for the train drivers and therefore they might not be very alert or differences are noticed quickly because it is not normal and they become very alert. For the same reason, the train drivers who will take part in the simulator test should have different driving experience (in time).

### 1.2 Problem statement

The problem we are facing is how ProRail can measure situation awareness in such a way that these measurements are useful for the safety officers (ProRail management) and key safety advisors (like RVT) to use in their decisions about implementing changes in signalling and other information to train drivers. The results should also be usable for other (future) projects so that strategic planners, who think about feasibility of change trajectories over a longer time span, can also use the results.

A lot about measuring SA can be found in the literature, but a single answer on what the best method is to measure SA cannot be found since this is dependent on the requirements. As can be read in Chapter 1.1 there are a lot of difficulties with measuring SA. The main problem that should be answered is: How to measure SA of train drivers taking into account that management should be able to make decisions based on these measurements.

### 1.3 Research goals

ProRail is contemplating applying SA measurements in the safety evaluation of changes to the current system. They want to know how they can measure SA and how they use the results to make decisions on it. Preferably, the results should also be (partly) usable in different settings. At the end of the thesis an appropriate method should be chosen, based on the requirements that are set in Chapter 1.4. The conclusion of the thesis should present a recommendation on how to design...
measurements of SA in the future in such a way that both project and strategic management can draw conclusions out of it.

1.4 Research questions

From the research problem the main research question and sub questions can be formulated. ProRail wants to measure SA of train drivers in the DSSU project because previous research suggested that the performance and behaviour of train drivers might be influenced. SA is used to compare the safety of the new situation with the current situation. Measurements for project and strategic management often have different purposes. Project management has an interest in the results primarily for their particular project, strategic management wants to know if the results also apply to other settings. Therefore the main research question for the thesis will be:

**How to measure SA of train drivers in order to compare the safety of different infrastructure designs, taking into account that both project and strategic management should be able to make decisions based on these measurements?**

To be able to answer this question, first some sub questions should be answered. To formulate these sub questions some requirements of the measurements are set. These requirements were set on forehand and based on literature and discussion with SA experts and ProRail employees. The requirements are divided into four categories: general SA measurement requirements, data analysing requirements, managerial requirements, and practical requirements.

**General SA measurement requirements**

First of all, SA measurements have some general requirements. These requirements are applicable to all SA measurements, independently of the area of application.

SA cannot be determined exactly: a single unit is not available; every method has its own unit and way of measuring. Also, an absolute ranking is not available. Therefore the SA of different designs will be compared. In order to be able to compare the SA of the different designs, the measurement method should be able to measure significant differences. A measurement is reliable when it can be repeated and will give the same values. The validity is also very important; the method should measure SA, not something else (Endsley, 1995a). Stanton and Young (1999) compare the difference between reliability and validity with gun shots. Reliability is the extent that shots are grouped, validity is the closeness of the shots to the target. Also when the measurements are taken, the SA of the driver should not be influenced (either positive or negative). Sometimes by asking questions about the SA of a person, his attention might be directed to something he was not thinking of before; in this case the SA of the person is influenced and biased data is measured (Salmon et al., 2006).

Recapitulating; the general requirements are:

- Sensitive enough to see differences between designs
- Reliable method
- Validated method
- Not influencing the train driver’s SA
**Data analysing requirements**

Second, there are some data processing requirements.

By making the SA measurements relatable to the GDTA and the SA-levels, the impact of the values can quickly be found. The GDTA is used to make a SA requirement analysis; all relevant tasks for operating a system are described in the GDTA. Endsley (2003) described three SA Levels: Level 1 perception of the elements in the environment, Level 2 comprehension of the meaning, Level 3 projection of future status. For example, if a low value is found on a specific SA question it gives an insight into which task is influenced by this SA value and what the reason of this value was (perception, comprehension or projection). If this is known it is easier to see the impact of this value on the safety and how it can be improved.

Since ProRail works with the Common Safety Method (CSM), it should be investigated if the CSM needs extra requirements for the measurements. In the CSM is stated that the safety performance should at least be the same as before. “This may require also explicit risk estimation in order to show that the level of risk is at least as good as that of the reference system” (European Railway Agency, 2009b, p. 42). The method should therefore at least use a reference system to compare the values with each other. Furthermore, the CSM has two main requirements for the explicit risk estimation and evaluation, in article 2.5.7. is stated: “The explicit risk estimation and evaluation shall satisfy at least the following requirements: (a) the methods used for explicit risk estimation shall reflect correctly the system under assessment and its parameters (including all operational modes); (b) the results shall be sufficiently accurate to serve as robust decision support, i.e. minor changes in input assumptions or prerequisites shall not result in significantly different requirements.” (European Railway Agency, 2009b, p. 45)

Recapitulating; the data analysing requirements are:

- Easily relatable to aspects of GDTA, so subsequently judge if measurement still has a value in changed tasks composition
- Measurement result should relate to the three SA levels
- The SA measurements should be applicable to the CSM method

**Managerial requirements**

Since the project- and process managers should make decisions based on the measurements, they should also accept the results.

It should be investigated what factors will influence the acceptability of the results. Because the measurements will be used to make decisions on safety, the relation should be very clear. What does the outcome of the measurements mean on the safety of the project.

It would also be preferable if the answers of the measurement are not multi-interpretable, the outcome should ideally give a single safety implication with one or more causes. These causes should not also be the reason of the contrary of the safety implication. If this is the case, it should be explained if the overall outcome will be negative or positive for the safety. For example, when the train dispatcher calls a train driver because the track is damaged 5 km ahead, the train driver gets a
better SA and will temper his speed. But the train driver is also distracted by the call and might therefore miss other dangerous situations during the call; he gets a worse SA. It should then be explained if the overall SA would be better or worse.

Finally, the measurements should be “future proof”. It is preferred that when in the future significant changes occur, the measurements are still useful. If this is not the case, it should clearly be stated what the conditions are in which the outcome is applicable.

Recapitulating; the managerial requirements are:

- Project- and process managers should accept the measurement method and procedure
- Measure results should be understandable in relation with safety
- Outcomes of the measurement should not be multi-interpretable
- SA values should be “future proof”; still useful in the future if other changes happen

**Practical requirements**

Finally, there are some practical requirements for measuring SA. These requirements specifically apply to measuring SA for ProRail and NS Dutch Railways.

Not all measurement methods are appropriate for measuring SA in certain circumstances (Salmon et al., 2006); the chosen method should be usable for train drivers and in a simulator. If ProRail wants to be able to do more SA measurements in the future, it should be executed without adjusting the simulator permanently. The simulators are used for training the train drivers as well; the trainers require that the simulators remain in an unchanged condition.

Ideally, the new infrastructure designs and processes are tested on the real track, but because of safety and practical reasons new situations cannot be tested on the real track. When the project is completed a monitor program will be used to monitor the drivers. Therefore the measurements have to be done in simulators. SA is influenced by all the elements in the environment. When not all the elements are present, the SA cannot be tested properly. The SA will then only be focused on the elements that are present; distraction elements will not be present. The effectiveness of evaluating designs is strongly influenced by the quality and realism of the simulations (Endsley et al., 2003). The handling of the train is also very important to be as accurate as possible; this can otherwise be a distraction that is not present in the real situation.

The training time for the measurements should be short for the train drivers if a lot of train drivers will participate in the study. Instructions are often needed for SA measurements; this in itself is not a problem. Train drivers are often only a day available for the measurements; long training time will then reduce the actual SA measurement time.

If a lot of measurements will be taken, it is likely that a lot of test instructors are needed. By making the test as easy as possible, the test instructors receive a simple training and will be replaceable.

Recapitulating; the practical requirements are:

- Useable for measuring SA of train drivers in simulators
- Simulator should not be adjusted permanently
• Simulator should be comparable with real situation
• Minimal training time for train drivers
• Easy to take tests for test instructors

Based on the main research question and the requirements several sub questions can be asked.

1. What is SA and how is it related to safety?
2. What are the requirements of decision-makers to accept the results of an SA study for the project?
3. What are the requirements for an SA study design for train drivers?
4. Which SA measurement method or set of methods is best to choose based on the requirements?
5. Are SA measurements useable to make a safety decision on?
6. How can the results of the SA study be generalized?

1.5 Research Methods and Outline of the report

Characteristics of the research methods chosen
In order to answer the research questions, two different research methods will be used: a desk research and a case study. A desk research is used to describe the existing theory in order to see the status of the research about SA measurements. A desk research is the best method to use for the literature study; theory can easily be accessed. An advantage of using desk research is that scientific information is widely available in books and articles. Also the information is easily process able. Using scientific articles that are published in a scientific journal has an advantage that scientific staff checked the content; e.g. the information is very reliable. The disadvantages are that there is no direct contact with the authors of the material. Also all the material used should already been published. Since writing books and articles take a lot of time, the latest update of the material is not always available. When companies have this data available it can be confidential information, which makes it hard or impossible to get.

The advantage of using a case study is that you can ‘test’ if the found answers are right, there might be things missing or wrong. It gives an insight in what outcomes of the SA-measurement can be obtained in practice. With these case studies, mistakes can be fixed before the report is published. Case studies are an appropriate method for studying behaviour and rare phenomena, since the researcher can study the subjects very closely. The disadvantage is that it can be very time consuming. Also not all variables can be chosen, since it is a case study some variables will be pre-determined. It will be hard to generalize the results and a possible bias in the outcome can occur since a single person analyses the information.

In Table 1 the overview of the chapters with the corresponding research questions can be found.
Table 1: Overview of research questions and their chapter numbers

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Method</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Literature study</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>Literature study</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>Literature study</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>Literature study + input from question 2,3</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>Case study DSSU</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>Literature study + outcome from question 6</td>
<td>6</td>
</tr>
</tbody>
</table>

Overview of the research methods per chapter

In chapter 2 the first research question will be answered: “What is SA and how is it related to safety?” A definition of SA will be chosen, which will be used in this thesis. Then the discussion about how SA is related to safety will be reviewed. Based on articles this first research question will be answered. Both the definition of SA and the discussion of the relation with safety is thoroughly described in literature.

The second research question will be answered in chapter 3: “What are the requirements of decision-makers to accept the results of an SA study?” First the influence of the safety culture on the decision-making will be investigated. Then the strategic decision making will be discussed. In this part will be discussed what the characteristics are from project decision-making processes. Finally the CSM will be discussed, since this risk method should be used by ProRail. All these questions will be based on literature, since these topics are extensively described.

Chapter 4 is a combination of literature study and the outcome from chapter 3. It answers the third research question: “What are the requirements for a SA study design?” First general study design literature will be discussed. Then based on the outcome of chapter 3 the study design can be set.

In chapter 5 the research question will be answered: “Which SA measurement method is best to choose based on the requirements?” Based on the study design a first selection can be made on which methods might be suitable. Then a literature study of the preselected methods will answer this research question.

When there is an advice about how the SA should be measured, it can be applied to a case study in chapter 6. This case study is about the Dutch DSSU project from ProRail, which will be explained in this chapter.

In chapter 7 the last research question will be answered: “How can the results of an SA study be generalized?” Literature describes three different theories about how qualitative research can be generalized.

The conclusion of the report will be given in chapter 8 based on the conclusions all the research questions.

The structure of the report follows the chronology of the project as given in Figure 1. First a specific project was executed, with all its steps: The DSSU project. This is displayed as the inner circle in the figure. The project goal is highlighted, since in this chapter the goal of measuring SA was explained.

This figure will appear in every chapter to highlight to which specific part of the project the chapter
contributes. Thereafter, when ProRail got the results of the DSSU project, they wanted to know if they could apply DDRR everywhere in the Netherlands. Therefore a system change should be made to allow DDRR everywhere, the results had to be generalized. This corresponds to the outer circle in the figure.

In Figure 2 the flow diagram of the thesis can be found. Every block in this figure represents a chapter. In the blocks is displayed to which part of the project the chapter contributes. This gives a visual overview of how the conclusion of this report will be made.

![Figure 1: Structure of the project](image)

1.6 Scope

This thesis will try to answer the question how SA measurements should be executed in such a way that decisions can be made upon the outcome by project- and in strategic management. General requirements of SA measurements and decision-making will be given. Examples will be given from the SA of train drivers. Then this is applied for a specific case study from ProRail in the Dutch railway industry.
Figure 2: Structure of the report
Decision-making based on measurements of behavioural safety
2 Situation Awareness

This chapter contributes to the problem definition of the DSSU project. A thorough understanding of situation awareness (SA) is needed before the theories of decision-making, study design, and measurement methods are discussed, in order to put these chapters into perspective. In this chapter the SA theory will be described. The three-level model of SA theory of Endsley will be used since this is the most developed theory (Stanton et al., 2001). Next the process of gaining SA via the mental models is explained, the function of schemata, and how this relates to expectations and automaticity. Finally the relation with safety is discussed. There is still a discussion going on about the relation of SA with safety, based on the different SA theories.

Figure 3: This chapter contributes to the project problem definition

2.1 Definition

Simply said, SA can be explained as the knowledge and understanding of a person about a situation in a certain place and time. Multiple formal definitions of SA exist. In this thesis the definition of Endsley will be used, since this is the most used definition in the literature. Endsley formulates SA as follows:

‘the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future’

(Endsley, 1988a, p. 97)
This definition exists of three levels (Endsley et al., 2003):

- Level 1: perception of the elements in the environment
- Level 2: comprehension of their meaning
- Level 3: projection of future status

In Figure 4 the three levels of SA are schematically drawn. All three levels will be discussed in detail.

### 2.1.1 Level 1: Perception

Perceiving the status, attributes and dynamics of the elements in an environment is the first level of SA. Important elements for train drivers can be: position of the train, signal lights, other trains etcetera. A train driver can for example see a yellow signal sign. All the important elements should be known to obtain a complete SA. All senses can be used to obtain the perception of elements, such as visual and tactile senses. For example a train driver can hear if the engine is working properly and see what the allowed speed is. Electronic displays and sound systems are often used to enhance the perceiving of important elements. Also communication with others can help to enhance the perception.

The confidence in the information perceived is very important; the information sources should be reliable. Information might be neglected when there is no confidence.

Problems with level 1 can be that important information is not perceived or presented. Distraction, by relevant or irrelevant sources, is the main cause that information is not perceived (Endsley et al., 2003).

![Figure 4: 3 Levels of SA (Endsley et al., 2003)](image)

### 2.1.2 Level 2: Comprehension

The second level of SA is the understanding of the information that was perceived. All information should be understood in its context. Since a lot of data can be presented at the same time, the data should be combined, prioritized, and linked to the goals. When a train driver sees a yellow signal light
Decision making based on measurements of behavioural safety

(perccepion), the driver should know that when he passes that light the maximum allowed speed is 40km/h. Depending on the current speed (perception) he knows whether he has to brake or not. A developed mental model is crucial in order to develop a complete level two SA. Inexperienced people therefore have a disadvantage (Endsley et al., 2003).

2.1.3 Level 3: Projection

In level three the meaning of the information on the future is developed. Level three SA cannot be developed without comprehension of the current situation (level two). When a train driver approaches a yellow signal light, he should know that the next signal light might be red and has to brake to a halt before that light. Experience is very important for a complete level three SA. Experienced people know a lot of scenarios that can occur and are constantly thinking forward to avoid unwanted situations (Endsley et al., 2003).

2.2 Mental model

Both working memory and long-term memory is very important for creating SA. The working memory is short-term memory that is used to constantly update the SA. A structure for long-term memory is a mental model. Rouse and Morris define a mental model as follows (1986, p. 7): “mechanisms whereby humans are able to generate descriptions of system purpose and form, explanations of system functioning and observed system states, and prediction of future states”. A visualisation of the relation between the mental model and SA is given in Figure 5. Both semantic and system knowledge are included in mental models. System knowledge is based on how a system functions. Semantic knowledge is about factual information; it tries to answer what instead of how (Endsley et al., 2003). An example of system functioning is: knowing how to drive a train. Memorising the train station names is an example of semantic knowledge.

A schema is a state of a mental model. It can be used to quickly identify situations that have occurred before, in order to quickly react to this situation. Schemata are developed by experience, either by personal experience or by others people experience. Young inexperienced persons do not have developed schemata. They can have difficulties with SA; they need more processing time to understand what is happening and they are likely to have a full working memory (Endsley et al.,
2003). The mental model and schema is an important factor for expectations and automaticity as described in chapter 2.4.

2.3 Goals

SA can be driven by goals or by data. Casson (1983) described goal-driven processes as top-down information processing, in which the goals determine which elements in the environment are important. For example, a train driver wants to drive his train safely and therefore pays attention to whether the doors are closed, if there are obstacles at the tracks etcetera. Data-driven processing is known as bottom-up processing (Casson, 1983). In this situation the data catches the attention of a person, independently of his goals. Flashing icons or alert sounds are examples of data-driven processing (Endsley et al., 2003).

The active goal determines to which element the attention goes. When a train driver departs from a train station, he is not thinking about how a train should be let into the depot. But when a person is totally immersed in his goals, he might not be influenced by other incentives. This can be very bad for the SA. Therefore, being able to quickly switch between goal and data driven processing is very important. Goals should be prioritized, because people often have multiple goals. Important information might be neglected if the wrong goal is chosen (Endsley et al., 2003).

2.4 Expectations and Automaticity

In the DSSU project expectations and automaticity might have a big effect. Train drivers have certain expectations of how the signal lights will behave. When the signalling configuration changes the SA might be influenced.

Expectations can have positive and negative effects on SA. Expectations help to quickly focus on important elements, but it might also help to make errors (Endsley et al., 2003). For example, a train driver can quickly identify signs when he knows where they are positioned, he does not have to search for them. But he might miss a newly placed sign, because he does not expect this sign (Endsley et al., 2003).

Automaticity can also have positive and negative effects on SA. It can free up mental capacity that can be directed to other inputs. But information can also be missed because of automaticity (Endsley et al., 2003). For example, a train driver does not have to think about how the train works. Giving traction and braking goes effortlessly. But when a driver passes a yellow light, which is normally green, he might miss this. He expects that it is green and therefore does not look closely anymore.

2.5 Safety

The relation of SA and safety is still subject of discussion. SA is developed for military pilots, but also
other fields use SA. A high level of SA was found to be critical in the aviation (Endsley et al., 2003). The discussion is mainly about the definition of SA and its underlying theory (Stanton et al., 2001). Three main theories of SA are developed. Endsley’s three-level model as described in chapter 2.1 (Endsley, 1988a), the interactive sub-systems models from Bedny and Meister (1999) and the perceptual cycle from Smith and Hancock (1995).

**Interactive sub-systems**

The interactive sub-systems model is an information processing approach to SA in which the processes involved are dependent on the goals (Bedny and Meister, 1999). The theory is based on the orientational structure of activity in Figure 6. Activity can be divided in stages: orientational, executive, and evaluative. A subjective model of reality is formed in the orientational stage. Goals, motives, internal states, and the result of mental actions influence the orientation of a person. In the executive stage, the action to achieve the desired goal is executed. In the evaluative stage, feedback is given from the executive stage to see if the desired goals are met (Bedny and Meister, 1999).

The eight functional blocks in Figure 6 represents how the comprehension of the meaning of the situation is gained. Stanton et al. (2001) summarised the roles and inputs in Table 2. Stanton et al. (2001) summarised the working of the model as follows: In the first block the information gets a meaning by applying the conceptual model of block eight to it, the ‘image’ of the goal, and the orientation of the needed action in block five. The output of the meaning is then again the input for the ‘image’ of the goal. In function block three a person chooses relevant tasks and goals, based on his own preference by using block four and five. The motivation of the tasks is gained in block four. This again is influenced by how relevant the person thinks the tasks are and the criteria it used from block six. Block four also gives input to block five in which the performance and engagement with the work is assessed. This gives input for the criteria of evaluation in block six. The past experience is stored in block seven, which is updated by the evaluation of block six. The conceptual model in block eight is the representation of the world as we see it, this again is updated by the experience in block seven.

<table>
<thead>
<tr>
<th>Block</th>
<th>Function</th>
<th>Input Block</th>
<th>Summary of role</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Meaning</td>
<td>0,2,5,7</td>
<td>Interpretation of information from world</td>
</tr>
<tr>
<td>2</td>
<td>Image</td>
<td>1,4,5,8</td>
<td>Conceptual ‘image’ of information-task-goal</td>
</tr>
<tr>
<td>3</td>
<td>Conditions</td>
<td>4,5</td>
<td>Dynamic reflection of situation and task</td>
</tr>
<tr>
<td>4</td>
<td>Evaluation</td>
<td>3,6</td>
<td>Comparing motivation and performance</td>
</tr>
<tr>
<td>5</td>
<td>Performance</td>
<td>3,4</td>
<td>Interacting with the world</td>
</tr>
<tr>
<td>6</td>
<td>Criteria</td>
<td>4,5</td>
<td>Determining relevant criteria for evaluation</td>
</tr>
<tr>
<td>7</td>
<td>Experience</td>
<td>6</td>
<td>Modify experience to interpret new information</td>
</tr>
<tr>
<td>8</td>
<td>Model</td>
<td>7</td>
<td>Modify world model to interpret new information</td>
</tr>
</tbody>
</table>
Perceptual cycle
In the perceptual cycle the SA is a interaction of a person with the world (Smith and Hancock, 1995). Smith and Hancock define SA as follows: “adaptive, externally directed consciousness” (Smith and Hancock, 1995, p. 137).

SA is formed by the interaction between an agent and its environment. Neisser’s (1976) perceptual-action cycle is the framework for this interaction as given in Figure 7. The environment is the input for the knowledge of a person. The knowledge is modified by combining previous knowledge and new knowledge. This knowledge then directs an action. This action samples or changes the environment and closes the cycle. Smith and Hancock added the invariant in the middle, “it forms the linkage among information, knowledge, and action that produce competent behaviour” (Smith and Hancock, 1995, p. 141). In this theory SA is not a snapshot of the mental model, but a process of modifying knowledge (Smith and Hancock, 1995).

Figure 6: Interactive sub-system (Bedny and Meister, 1999)

Figure 7: Neisser’s (1976) perceptual cycle, with the invariant in the core (Smith and Hancock, 1995)
The three theories can be divided in two categories: SA as a product or as a process. The three-level model and the interactive sub-systems model see SA as a product (a state of the mind), whereas the perceptual cycle theory of Smith and Hancock see the SA as a process (interaction between environment and person) (Stanton et al., 2001). Both categories do see a link with safety. Endsley suggested that low SA leads to slower detection of problems and needed more time to investigate problems and conducting activities when a problem is detected (Endsley and Kiris, 1995). Smith and Hancock link SA to the capacity to generate consciousness: “We argue that SA is more than performance. More fundamentally, it is the capacity to direct consciousness to generate competent performance given a particular situation as it unfolds” (Smith and Hancock, 1995, p.138). Although the SA is not yet uniformly defined, lots of research suggests that lack of SA was the underlying problem of accidents. Hartel, Smith & Prince (1991) reviewed more than 200 aircraft accidents, they stated that SA was the main factor of these accidents. Sneddon, Mearns and Flin (2006) investigated 332 incidents in the offshore industry which 135 were marked as SA errors. In the Ladbroke train collision SA is also stated as the cause of the accident (Stanton and Walker, 2011). Some research even had the accidents categorized in level 1, 2, and 3 errors from Endsley’s model: in the aviation domain 76.3% of the errors were due to lack of level 1, 20.3% is due to error in level 2 and 3.4% is due to error in level 3 (Jones and Endsley, 1996). We can conclude that although the exact relation between safety and SA has still to be determined, a lack of SA is an important factor for accidents. Having a complete SA is needed for making an informed, conscious decision and therefore important for safe operations. But be aware, although SA is a requirement, it is not a guarantee for making informed, conscious decisions. More factors than SA have an influence on decisions, such as technical failures and skills. A student train driver might be very aware of a red sign, but not stop the train because he does not know how to do this.

2.6 Conclusion

Endsley’s three-level SA model is the most developed approach for the measures and intervention (Stanton et al., 2001). In Figure 8 a complete overview of SA in dynamic decision-making is given. The three levels are modelled in the SA; perception, comprehension, and projection. From this three levels the SA of a person is formed, which he uses to make a decision. Based on his decision the situation will change and then the SA cycle begins again. A lot of factors influence the SA and the decision of a person. The most important ones are the mental model, in which long-term memory helps to recognize the situation, goals and objectives, which help to focus on the important elements in the environment, and expectations and automaticity, which can both help and hinder for gaining SA.

The exact relation is yet to be set between safety and SA, but in the past a lot of accidents were already identified in which lack of SA was the cause of the accident. Also, all individual SA theories independently make a relation between SA and being able to perform well. In Figure 8 can be seen that SA is the input for decision-making. In other words, without a complete SA, an informed, conscious, and safe decision could not be made. But the contrary is not true, a safe decision is not guaranteed with a complete SA.
Decision-making based on measurements of behavioural safety

Figure 8: Model of situation awareness in dynamic decision making (Endsley, 1995b)
3 Decision-making

A part of the problem definition of the DSSU project is the acceptability of the SA results. To be able to get the results of this SA research accepted, it should be investigated how safety is integrated in decision-making. Today, many companies state that safety is their top priority, however this was not always the case. Historically, companies did not have any incentive to have a safety profile since it did not have any commercial advantages. Nowadays companies do have safety profiles but they have to trade off safety against costs. In contrast to private projects, big infrastructure projects are very susceptible to bad publicity (Whittington et al., 1992). The public often has an opinion about accidents or project failures of these companies, which can have bad consequences for the company. For this reason they spend a lot of attention to safety.

First, ProRail’s safety culture is described, according to the safety culture ladder. Negative study results might be accepted easier when safety is their top priority. If safety is not the top priority, commercial interest might get the upper hand in decisions. The second part explains some characteristics of strategic decision-making with their pitfalls. Decision-makers have three main stages of decision-making: goal formulation, alternatives generation, and evaluation and selection. Negative study results might not be easily accepted if safety is not incorporated in all three stages. The third part discusses the Common Safety Method. ProRail has to use this method for evaluating safety risks for new projects, therefore this is important for the formal acceptability of the results.

Figure 9: This chapter contributes to the project problem definition

3.1 Safety culture

Reason (2000) defined safety as: “the ability of individuals or organisations to deal with risks and hazards so as to avoid damage or losses and yet still achieve their goals” (Reason, 2000, p. 5). The
culture of the organisation can have a big impact in this ability. Organisational culture is a framework in which groups and individuals work which consists of attitudes of a national, organisational or professional values (Helmreich and Merritt, 1998). The safety culture are the beliefs specifically to health and safety within the organisational culture (Clarke, 1999). Parker et al. (2006, p. 552) summarizes Reason’s (1997) proposals for organisations to obtain an effective safety culture:

- “Has a safety information system that collects, analyses and disseminates information from incidents and near misses, as well as from regular proactive checks on the system;
- Has a reporting culture where people are prepared to report their errors, mistakes and violations;
- Has a culture of trust where people are encouraged and even rewarded to provide essential safety-related information, but also in which it is clear where the line between acceptable and unacceptable behaviour is drawn;
- Is flexible, in terms of the ability to reconfigure the organisational structure in the face of a dynamic and demanding task environment;
- Has the willingness and competence to draw the right conclusions from its safety system, and is willing to implement reform when it is required.”

Since the Dutch railway sector is divided in the rail infra structure manager (ProRail) and train operators (NS Dutch Railways, Arriva, Veolia etcetera), ProRail has to corporate closely together with the train operators to obtain a high safety culture. Especially the communication between the two companies should be well-organized: if NS drivers communicate problems with their bosses, but these bosses do not communicate with ProRail, it is still very difficult for ProRail to act on these problems. The other way around holds the same: ProRail should communicate with NS Dutch railways in order to let them to be able to act on safety hazards.

According to (Westrum, 1988) a companies safety culture can be classified in three different categories: Companies with a pathological style of management only meet external set goals and should be controlled and enforced. Companies with a calculative style of management tend to focus upon individuals and concrete technical issues. Safety issues are only taken into account if there is proof that the improvements work. Companies with a generative style of management respond positively to suggestions and information. The company itself is the main drive for the safety improvement (Whittington et al., 1992). Parker et al. (2006) added two additional levels, which were initially proposed by Reason (1997): reactive and proactive cultures. In a reactive style a lot of investigation is done about safety after an accident has occurred, a proactive style tries to mitigate safety hazards before they arise. In order of non-developed to well-developed safety cultures the total list as proposed by Parker et al. is:

- Pathological
- Reactive
- Calculative
- Proactive
- Generative

In a discussion held with Groeneweg (2013), safety advisor for the DSSU project and experienced safety advisor in the oil and gas industry, ProRail’s safety culture was rated as quite low. There are
some people in ProRail who have a proactive attitude towards safety, but safety issues are sometimes too easily accepted. In another discussion with Ransijn (2013a), safety advisor for ProRail, ProRail was rated to have a high safety culture, especially compared to other rail companies in the Netherlands. ProRail itself stated that safety is their absolute priority; therefore they try to improve the safety culture (ProRail, 2013a).

One of the aims of the government, shareholder of ProRail, is to make the rail safer. This should be demonstrated by diminishing the signals passed at danger (SPAD), less level crossings incidents, better safety organisation of construction outside, always knowing what the content of a wagon is and weighed routing of trains with dangerous substances (Ministry of Infrastructure and the Environment, 2013). The government does not give any target numbers in this document. ProRail has set itself the goal to have zero avoidable accidents (ProRail, 2012), which is one of the characteristics of a calculative style of management (Reason, 2000). These companies are trying to have a ‘safety war’ with belief that they can win this war entirely. A dangerous misinterpretation in this is that this war can never be defeated, but companies might think that they have won (Reason, 2000). When hazards are found an organization can only defend against it, it cannot eliminate them. An organization can only seek to minimize the incidence of unsafe behaviour, it cannot completely eliminate the basic human propensities for emitting errors and violations (Whittington et al., 1992).

High reliability organisations (HRO) take a different approach; they see this ‘safety war’ as an endless guerrilla conflict. HROs try to work within the safety limits to achieve their production goals; “they anticipate the worst and equip themselves to cope with it” (Reason, 2000, p. 11).

Some examples of a reactive safety culture at ProRail can be found. After a severe collision in Amsterdam on 21 May 2004, several measures were taken to lower the amount of SPADs (van den Top, 2010). Four projects started: Analysis of station layouts, a drivers program was set-up, route setting instructions to prevent automatic use of routes which had regular SPADs, a patch to the ATB-system (Schmeink and Beuk, 2006). After another severe collision in Amsterdam on 21 April 2012, a report was made up with recommendations for NS Dutch railways, ProRail, and the ministry of Infrastructure and environment. Some of the comments in the summary give indication for a low safety culture. For instance, the train dispatcher does not have any adequate devices to alert the driver that a red sign has been passed. Such devices did exist, but it was removed since it did not meet the requirements. A new system was not implemented afterwards. (Onderzoeksraad voor Veiligheid, 2012). In this particular example, ProRail was able to identify a risk that train dispatchers might miss a red signal passing, but did not mitigate the risk after the first system was rejected.

Companies should avoid reactive safety. In reactive safety, safety is implemented when something already went wrong. Companies that apply reactive safety often have a technical or hardware approach to safety and a rules and regulations safety culture. They have a traditional safety view of human performance and an emphasis on short-term solutions to safety problems rather than on an attempt to identify more deep-rooted organizational failures. Rather than responding to individual accidents, companies need to constantly reviewing the extent to which company procedures deliver the appropriate standards and identifying those situations in which these standards become compromised (Whittington et al., 1992).

Before systems are implemented they should first be tested, this can be done via gaming simulations. Testing via gaming simulations give the opportunity to mitigate hazards before a project is executed.
ProRail used to test a project before it was taken in use but after it was already executed. When gaming simulations are used, more time is available to solve the problems. Technical problems are solved with technical implementations, procedural problems with procedural implementations. When systems are not tested (long) before the actual implementation, technical problems were also solved with procedural solutions, and procedural problems were also solved with technical solutions, the problem was not changed in its core (van den Hoogen and Meijer, 2013). By measuring SA before a project is executed, safety issues from human errors can be acknowledged and treated. This is a proactive attitude towards safety.

Some companies approach safety through a set of rules. At ProRail, new designs are implemented according the OVS documents (in Dutch: ontwerpvoorschrift, English: design requirements documents). A problem is that safety departments might be considered as necessary but they are not very influential in an organization since safety is difficult to link to the commercial success of a company. To make sure that safety is more important in the decision-making, managers should be held accountable for the safety. When managers are accountable for safety they will feel more responsible. If they are not accountable for safety they will most likely focus on other things that are important for the management such as production, quality and cost (Whittington et al., 1992). This is already implemented at ProRail; line managers are responsible for the safety (Ransijn, 2013a).

Including safety can result in higher costs for realizing the project or lower production rates. The decision-makers have to weigh the consequences of the risks with the costs of the corresponding mitigation. Safety performance is dependent on management decision-making and the existence of effective management controls. To achieve a high safety culture in the company the safety goals should be implemented with other company goals. An integrated approach to safety management is needed (Whittington et al., 1992). Also the safety behaviour of the workers are formed by their perceptions of managers' behaviour and attitude in relation to safety (Clarke, 1999). A positive safety attitude of senior management is even considered as essential for a positive safety culture (Griffiths, 1985).

Overall, ProRail cannot be rated to have a high safety culture. Most references in this thesis indicate a middle or low safety culture. This chapter is too short to give a definite value to the safety culture of ProRail, for a definite answer a thorough audit is needed. The limited review in this chapter indicates a reactive or a calculative culture. The company is not neglecting safety issues as a whole, but it does not always succeed in mitigating risks. They are trying to come higher on the safety culture ladder.

3.2 Strategic decision-making

Besides safety culture, strategic decision-making is very important for the acceptability of the results. In the end, decision-makers are the people who have to accept the results of the study. Decision-makers have to make strategic decisions before a project is executed. This is executed in three steps: goal formulation, alternatives generation, and evaluation and selection (Schwenk, 1984). Schwenk (1984) describes some pitfalls of all these steps. In a goal formulation or problem identification, the problem is set with its corresponding gap. A gap should be recognized in this stage by the decision-
makers between the desired situation and the current situation (Schwenk, 1984). If there is no consensus between the decision-makers about the goal formulation or the problem identification, the results might only be accepted if it is positive for their own (hidden) agendas. In goal formulation four problems might occur.

Firstly, decision-makers can have a prior hypothesis bias. When people form a wrong hypothesis about variables they are not likely to change their beliefs even when they have evidence that they are wrong (Levine, 1971, Pruitt, 1961, Wason, 1960). This corresponds with Reason’s statement: the certainty of outcomes is very important (Reason, 1990). If decision-makers can doubt about the certainty of safety outcomes, they will even be more likely to neglect the outcome. The problem with safety is that the outcomes are of a probabilistic character this is in contrast to the outcomes of production, these are always very certain. Also the nature of feedback between production and safety is different. Production goals are unambiguous, rapid persuasive and reinforcing. Safety goals are negative, intermittent, deceptive and only persuasive after a major accident. Decision-makers do not always interpret information accurately when feedback is given about safety. Managers are often defensive on this feedback (Reason, 1990).

Secondly, decision makers can have difficulties with adjustment and anchoring. Decision-makers should make initial judgments and adjust their judgments when new information becomes available. The adjustments made are often insufficient (Tversky and Kahneman, 1974).

Thirdly, decision-makers can have escalating commitment. When they commit resources to a project, they are likely to defend that project. They can feel personal responsibility for the project and therefore keep investing in the current project even when there is evidence that the strategy is not paying off (Staw, 1981). When the time between the start of the project and the feedback of the results is getting longer, it is more difficult to stop the project, since more resources are committed. Groeneweg suggested that ProRail should incorporate the stage-gate model (Groeneweg, 2013). “A Stage-Gate process is a conceptual and operational map for moving new product projects from idea to launch and beyond—a blueprint for managing the new product development (NPD) process to improve effectiveness and efficiency” (Cooper, 2008, p. 214). A decision gate should be passed after every stage of the design and implementation process. By applying this model the loops of decision-making are very small, this makes it easier to adjust or stop a project.

Fourthly, decision-makers might reason by analogy. When analogies of simple situations are used for complex strategic problems it can help the uncertainty of other actors. The problem with the use of analogies can be that the situation can be sketched to simplistic (Steinbruner, 1974). If this is the case, the decision-makers are incapable of making a well-informed decision.

When the goal formulation or problem identification is done, the alternatives generation has to be executed. Often multiple alternatives can be developed for a problem. Possible appropriate alternatives might be left out when a problem definition is chosen to narrowly (Alexander, 1979), which makes it impossible to choose the best alternative. According to Schwenk (1984), there are four reasons why this occurs.
Firstly, decision-makers may focus too much on one of the goals instead of focusing on multiple goals (Steinbruner, 1974). Therefore they might focus on a single alternative and neglect others that are scoring better on other goals.

Secondly, decision-makers may search for negative aspects of other alternatives to convince themselves that these alternatives cannot be implemented (Steinbruner, 1974). The alternative that they prefer will then seem to be better than other alternatives.

Thirdly, facts can be interpreted in such a way that their own alternative seems to have benefits on all values. It appears that no negative characteristics off their own alternative is present (Steinbruner, 1974, Jervis, 1976).

Fourthly, it is very hard to develop alternative strategies when often one problem-solving strategy is used (Anderson and Johnson, 1966). Often multiple problems are solved in a new design. Multiple problems combined can have multiple solutions.

When multiple alternatives are available, these should be evaluated. Schwenk (1984) gives three possible problems with the selection of alternatives.

Firstly, there might be a representative bias. A decision-maker might overestimate the representativeness of a situation in which he wishes to generalize. This involves the insensitivity to predictability of the effects (Tversky and Kahneman, 1974).

Secondly, an illusion of control might appear. Decision-makers might think that outcomes of a strategy are under their control. They think that when problems arise they can solve the problems themselves (Lefcourt, 1973, Langer, 1975, Larwood and Whittaker, 1977).

Thirdly, partially described alternatives might be devaluated because it involves uncertainty (Yates et al., 1978). These alternatives might be the best to choose, but because of the uncertainty they are not chosen. Full description of an alternative is therefore very important.

Safety should be incorporated in all the steps to make sure that the study results are accepted. A company with a high safety culture will have safety incorporated in all these steps. When the safety study of the DSSU project started, it was already decided that DSSU was the only option for adding capacity. No alternatives were considered at this point. Therefore ProRail had a big interest that the results were positive, they stated that if the results will be negative that they will accept this and stop or adjust the DSSU project.

In order to enhance the acceptability of the results in the DSSU project, the decision-makers were kept closely informed on the process. Every time uncertainties or decisions on study designs should be made, decision-makers had to decide on this in order to make it ‘impossible’ for them to reject the results. The consequences of the decisions were explained for the robustness and the applicability of the results.
3.3 Common safety method

A formal requirement of the acceptability of the results is the use of the Common Safety Method (CSM). In the rail sector every introduced change should be checked if the CSM is required. This method is developed by the European Railway Agency and must be used by all rail companies in Europe (European Railway Agency, 2009a). Therefore the CSM is used for the DSSU project (Ransijn, 2013b). In Figure 10 the flow chart for determining if the CSM is required is given. All the questions should be answered by expert judgment. First should be determined if the change is relevant for the safety. If this is not the case the CSM does not have to be used. Extra questions should be answered to judge if the change is significant enough to apply the CSM when the change is related to safety. The following questions should be answered separately:

- Is there a low failure consequence?
- Is the novelty low?
- Is the complexity of the change low?
- Is easy monitoring possible?
- Is it easy to revert the system to previous state?
- Is the change still insignificant if all recent safety-related changes are taken into account that were not rated as significant?

All answers to the questions should be documented. The proposer can judge, based on one or more of the questions, if the change is significant. If all the questions are answered positively the change is not significant and the CSM does not have to be applied (European Railway Agency, 2009b).

3.3.1 System definition

The CSM flow diagram can be found in Figure 11. If the CSM is required, the system definition should be made. In this definition the system objective, system functions, system boundary, the physical and functional interfaces, existing safety measures and safety requirement and the assumptions that shall determine the limits for the risk assessment is given. The system objective should give the intended purpose of the proposed change. The system functions and elements should contain all relevant functions of human, technical and operational elements. The system boundary should contain to what extent the system changes including other interacting systems. The system environment should contain all important characteristics in the environment such as shocks, vibrations and operational use (European Railway Agency, 2009b).

3.3.2 Hazard identification

Next the hazard identification (HAZID) and classification should be made with the help of wide-ranging expertise form a competent team. ProRail already has a generic hazard log. Every new hazard found in the new system description should be added to the generic hazard log to prevent that hazards of previous system description that were neglected are now again analysed (Ransijn, 2013b). If the risks are acceptable the CSM can be ended. If the codes of practice or a reference system is used to control the risk, the HAZID should also contain the verification and deviation of the proposed change in comparison to these methods (European Railway Agency, 2009b).
3.3.3 Risk acceptance strategy

When the risks are not acceptable, a risk acceptance principle should be chosen. Three options are available: codes of practice, similar reference systems, and explicit risk estimation.

3.3.3.1 Codes of practice

ProRail has a lot of regulations for the implementation of new techniques and design criteria in the OVS. If the new design is within the boundaries of these regulations the codes of practice can be used (Ransijn, 2013b). The codes of practice should be widely recognized in the railway domain, it must be relevant for the control of the hazards, and it should be available for assessment bodies (European Railway Agency, 2009b).

3.3.3.2 Reference system

When the codes of practice cannot be used a reference system can be searched. This reference should have similar functions and interfaces, similar operational conditions, similar environmental conditions and should already been proved in-use to have an acceptable safety level. If all these requirements are met, the risk of the reference system shall be considered as acceptable and safety requirements may be derived from the reference system. If the system deviates of the reference system, it should be demonstrated that the safety level is at least as good as the reference system (European Railway Agency, 2009b).

3.3.3.3 Explicit risk estimation

If there is not a codes of practice or a reference system available an explicit risk estimation should be made. This can be done quantitative or qualitative. An explicit risk estimation was made for the DSSU project. To accept the risks of a project the safety performance should at least be the same as before. The method should therefore at least use a reference system to compare the values with each other. The methods used should reflect the system correctly under assessment and its parameters and the
results must be sufficiently accurate to serve as a robust decision support (European Railway Agency, 2009b).

### 3.3.4 Risk evaluation

When the risk estimation is executed the risks must be evaluated based on the legal requirements of the community legislation or in notified national rules. Technical failures should have a failure rate of maximum of $10^9$ per hour. If the risks are considered as acceptable, the safety measures should be described in the hazard record. If the risks are not acceptable, additional safety measures should be implemented.

### 3.3.5 Safety requirements

Additional mitigating factors may be set after the risk evaluation stage, if the risks are not acceptable according to the safety requirements as set in the system definition. If the risks are acceptable the proof should be given that the safety requirements are set.

### 3.3.6 Demonstration of meeting safety requirements

When the risks are evaluated and the safety requirements are set it should be demonstrated that the safety requirements of the new system is met. The proposer should demonstrate this to an independently assessed body. If the safety requirements are not met, mitigating factors can be identified, and the system definition can be updated in which the CSM should be run again from the beginning until the residual risk is acceptable.

If the risks are acceptable a monitor program can be started to monitor the safety. When the monitor program is finished, the system description can be added to the codes of practice (Ransijn, 2013b).

The entire risk assessment process should be documented and include (Barrosso, 2013, p. 22):

a) “description of the organisation and the experts appointed to carry out the risk assessment process;

b) results of the different phases of the risk assessment and a list of all the necessary safety requirements to be fulfilled in order to control the risk to an acceptable level;

c) evidence of compliance with all the necessary safety requirements;

d) all assumptions relevant for system integration, operation or maintenance, which were made during system definition, design and risk assessment.”

### 3.4 Conclusion

In this chapter, three factors for the acceptability of the SA study results are discussed: Safety culture, strategic decision-making and the CSM.

Although this chapter is to limit to give a definite answer, it is indicated that ProRail is not on a high level of the safety culture ladder, examples of ProRail's reactive and calculative culture are found. ProRail is not neglecting safety issues, but it is not the most important decision criteria. New designs are often rated according to a set of rules. It is likely that negative study results will be accepted if the outcome is clearly very bad. When negative study results are not clearly very bad, other aspects
might become more important in taking a decision about the project such as economic reasons. The DSSU project is first tested on safety issues before it is implemented with SA measurements among other measurements. This is a first step in a more proactive safety culture.

Safety should be incorporated in all three levels of decision-making: goal formulation, alternative generation, and evaluation and selection. A big risk of the DSSU project is that no alternatives were left when the safety study was executed. Stopping the project was at this point very hard. ProRail stated that if the results are negative they will accept this and stop or adjust the DSSU project. Again, there is a real danger when the results are negative but not significant. If the gate model was used in the execution of the project, every gate could be tested on safety. By using this model it is easier to adjust or stop the project, based on safety measurements outcomes. The loops of decision-making are short, which decreases the escalating commitment of the decision-makers. For enhancing the acceptability of the results the decision-makers were kept closely informed on the status of the DSSU project. They had to decide on all the decisions that had an impact on the uncertainties and applicability of the results, which makes it very difficult for them to neglect the study results.

The CSM is used by ProRail for determining the risk of the DSSU project. The formal requirements for the decision-making are: the safety should be at least as good as reference systems, the methods used should reflect the system correctly under assessment and its parameters and the results shall be sufficiently accurate to serve as robust decision support.
Figure 11: CSM scheme (European Railway Agency, 2009b)
Decision-making based on measurements of behavioural safety

René van der Meij
4 Study design

In the previous chapters the problem definition is described. The acceptability of studies gives some requirements to the study design. A rigid design is needed to face these problems. In this chapter an overview is given of different SA measurement designs with their corresponding characteristics, which contributes to the project experimental set-up. First some general study designs are given, such as before and after studies. Second a short description is given for measuring human performance; humans can behave differently when they know they are tested. Next the measurement points are discussed, describing what and at which moment should be measured. Finally a mixed method design is discussed; outcomes of mixed method designs are more rigid. As discussed in chapter 3, rigid outcomes are very important for the acceptability of the results.

![Figure 12: this chapter contributes to the project experimental set-up](image)

4.1 General study designs

Study designs can be categorised in controlled or uncontrolled studies. In a controlled study the researcher can control one or more aspects of the study. A controlled study can be analysed with a before and after study, uncontrolled can only be analysed with observations after an intervention took place. When there are multiple controls possible a quality control, controlled before and after study and a stepped wedge design, is possible. If only the time can be controlled a before and after study is the only option to choose. Unfortunately, this is not the best method for distinguishing cause and effect since simultaneous changes will influence the outcome of the study (Brown et al., 2008a). In a before and after study, the measurement before an intervention is compared with a measurement after an intervention (Robson et al., 2001); in the SA study ‘before’ is used to research the current situation, ‘after’ is used to research a new design. Three recommendations can be made to increase the confidence in the outcome of the study (Brown et al., 2008a): First, multiple before
and after measurements will make the outcome statistically more grounded. If only one or two measurements are done, the outcome is highly influenced by a single (wrong) outcome. Second, a big magnitude and rate of change of the input will help to convince that this change is accountable for the change in the outcome. Third, the plausibility of the change should be logically linked to the change in outcome. If the evidence of the outcome is compatible with other evidence, the evidence is stronger (Brown et al., 2008a).

Controlled comparative studies can also be used instead of before and after studies, in which one group is exposed to the change of the study and the other group is not (Brown et al., 2008a). This method is very suitable when an intervention cannot be undone. Controlled comparative studies are like an experimental design, but the participants are not randomly assigned (Robson et al., 2001).

Designs can be non-experimental, such as a natural experiment, quasi experimental, or experimental, in which interventions are chosen randomly (Brown et al., 2008a). The most unreliable experiment design will be the non-experimental design with only after observations, in this study the interventions cannot be controlled and no comparison is possible of the data before and after the intervention. Clearly, the experimental design with before and after measurements is the most reliable one, the interventions are controlled and a comparison can be made of the data before and after the measurement (Brown et al., 2008a).

In a stepped wedge design the subjects get an intervention at a different moment, but the intervention is not withdrawn anymore when it is applied to the subject. More data collection can be done in this design, all the data from the before intervention can be compared with the data after the intervention (Brown et al., 2008a). This method is very useable for medicine research.

A comparison can be made between clusters or between individuals. A bigger sample is needed when comparisons are made between clusters, since there is a loss of power. When using clusters the similarity of individuals within the cluster will be greater than the similarity across the cluster. The sample size depends on the size of the effect and the risk of false study results that are within the acceptance limit (Brown et al., 2008a). Two types of cluster designs can be distinguished: cohort designs and repeated cross sectional designs (Ukoumunne et al., 1999). In a cohort design subjects get repeated measures. In repeated cross-sectional designs a sample of subjects gets measured in every measurement. A cross sectional design is specifically suited for measurements with terminal end points, in which an end point can only occur once (Brown et al., 2008a).

Outcomes of a study can be influenced by a lot of factors, also factors that were not included in the study. Therefore a big signal to noise ratio should be achieved in which the signal is the intervention and the noise is the variance in outcomes due to other factors. To improve the signal to noise ratio the end points should be chosen which are exclusively an outcome of the intervention. A second improvement can be to choose only end points that were caused because the intervention did not take place. A problem is that the outcome can be very biased because researchers select the wrong endpoints (Brown et al., 2008b).

Measurements should be consistent to avoid bias. Observers should have the same understanding of definitions. By spreading the observers over multiple sites, the bias error also diminishes. Case-mix can also contribute to bias. When on one site there are more subjects with a certain characteristic,
comparison of subjects can be biased. Randomisation and controlled before and after measurements can minimise this bias (Brown et al., 2008b).

Brown et al. (2008c) state that study design is influenced by four factors:

- Logistical/pragmatic constraints
- A priori assessment of benefit and harm
- Plausible effects on end points
- Target audience

The logistical and pragmatic constraints depend on the rarity of an event to happen and the timing of the introduction of the intervention. Quantitative designs are most feasible for events that are likely to occur. Statistics can be applied when end points occur often. When an event is not likely to occur, process end points, such as intervening end points, can be used. By using these end points, more data can be acquired. A before and after study can be used if an intervention is planned, otherwise a measurement before the intervention cannot be executed. Concurrent controls can be used when an intervention consists of multiple changes, unless these changes cannot be implemented independently of each other (Brown et al., 2008c).

An a priori assessment of a project can be necessary to investigate the risk, costs and benefits. A simple quantitative study will suffice when an intervention is likely to be low cost and not risky. For high cost and risky interventions a more extensive study is needed with before and after measurements in the control and intervention group. A Bayesian prior can help to explicitly state the effectiveness of an intervention. Experts should determine this probability distribution. By combining this data with direct data of the measurement, the probability can be updated. The cost effectiveness of the intervention can then be derived from the midpoint of this probability distribution (Brown et al., 2008c).

The type of study also depends on the effects on the end points. A small improvement on a common error will already have an effect. To study these improvements, a prospective comparative study can be used. Researchers can be biased when investigating small improvements, therefore contemporaneous controls should be used and observers should be masked. Improvements on rare errors should only be executed when the improvement is big, otherwise the intervention will not be cost effective. A simple before and after study will be sufficient to see the effects of the interventions (Brown et al., 2008c).

The target audience will also affect the choice of the study design. Different audiences will demand different levels of evidence. Local managers who are responsible for their area will mainly have local interventions; a simple before and after study will suffice. Managers who are responsible for multiple areas will need more detailed evidence, since the intervention will affect a large area. They will want to know a detailed relation between cause and effect and be able to generalise the results (Brown et al., 2008c).

Although detailed studies will sometimes be preferred, it is not always possible to execute them. Brown et al. (2008c, p. 180) states: “The distinction is sometimes captured in the language of sufficing (doing the best what is available) versus optimising (doing robust scientific studies that may shift international opinion and practice”).
4.2 Human factor study

In a study it is important that subjects do not know what points are being measured in a study, since they might adjust their behaviour (Brown et al., 2008a). Sometimes this can be achieved by misleading the subjects about what is being measured or in a medicine study by using placebo’s in a stepped wedge design. It is not always possible to hide what is being measured, for instance when the measurement method needs verbal or written input from the participant. In that case the moments of measurements can be chosen randomly. Participants cannot prepare for the measurements when they are unaware of the moments of measurement.

4.3 Measurement points

End points should always be valid and reliable. The validity of an endpoint is the accurate reflection of the intended underlying concept. In other words, it should measure what it states that it will measure. A reliable end point is repeatable; the same conclusions should be drawn when a study is executed for a second time. Reliable implicit reviews can be reached via training (Brown et al., 2008b).

During the implementation of the interventions some process evaluations can be made. First, the fidelity of the intervention can be measured; is the intervention the same as was intended? Intervening variables, such as knowledge or beliefs, can help to evaluate diffuse interventions. If these variables can be measured, the policies can already be monitored during the implementation process. Intervening variables can be used to research what the cause of the effect is, or what the effect of the cause is. With the former can be tested if the intervening variable indeed predict the outcome of the end points; this is called the validity of the surrogate. In the latter is known that the intervening variable predicts the outcome of the study, it is used to demonstrate the effectiveness of the intervention(Brown et al., 2008b).

Error is the final surrogate point. If errors occurred the wrong conclusions can be drawn from the study. Errors can be measured via reporting systems. The problem with reporting systems is that it can be incomplete and selective or the underlying problem is not described. Trigger tools can help to identify strange values in outcomes. It should be sensitive and specific enough to measure the outcomes. Implicit and explicit reviews can help to collect the errors. In an implicit review the expert makes a judgement based on their own judgements, in explicit review standards are used to identify errors. These standards are developed by a group of experts (Brown et al., 2008b).

4.4 Mixed method design

Combining the outcomes of different studies can make a study more robust; this is called a mixed method design. In a mixed method design both qualitative as quantitative studies are combined, each with their own end points. Mixed method designs have four main advantages (Brown et al., 2008c):
Decision-making based on measurements of behavioural safety

• Triangulation
• Understanding
• Development of theory
• Revisions of study protocols

Triangulation is combining the conclusions of different studies. Conclusions are more robust when it is confirmed by other findings. Triangulation also helps to the generalizability of the findings. When other findings confirm the conclusion it is safer to generalize the results. The danger of ‘over-generalization’ is diminished when other findings disclaim the conclusions (Brown et al., 2008c).

Different perspectives are gained on a conclusion when mixed method design is used. This enables the researchers to consider why the outcome is effective or not effective (Brown et al., 2008c). Also the underlying processes might become clear.

The outcome of one study can help to define research questions for other studies. Theories can be developed when these studies are also executed (Brown et al., 2008c).

Sometimes an intervention should be changed or discontinued because it is unsafe or an undesired effect occurs. When multiple studies are executed simultaneously, a problem can be observed early and it is easier to change the intervention (Brown et al., 2008c).

Combining methods can be used as a complementary study, in which the methods are closely linked together (Whitehead and Yin, 2003). One method can research the frequency of certain events, while the other method can research the underlying processes.

4.5 Conclusion

The best study design for measuring SA in a simulator is a controlled before and after study. By using a simulator, the researcher has the best control over the changed factors in a design and therefore make the result of the study a direct outcome of the introduced intervention. When a lot of participants are available, a cluster design should be used, otherwise an individual comparison should be chosen.

The participants should not be informed about what is measured, since people might show different behaviour when they are studied. If this is not possible measurement points should be chosen randomly, in such a way that the participant does not know when the measurements will occur so that he cannot prepare himself.

End points should be valid and reliable, it should measure what is stated that it will measure and should be repeatable. Surrogate points can be chosen if end points cannot be measured, in which the surrogate endpoint should be able to predict the outcome of the study.

Different methods can be combined in a mixed method design, in order to make the outcome more robust. When multiple methods give the same outcome, the certainty of the outcome is better. Mixed method design can also be used in which one method researches the frequency of an event, and the other method researched the underlying processes.
Decision-making based on measurements of behavioural safety
5 SA measurement methods

The last step of the experimental set-up of the project is the selection of the SA measurement method. In the previous chapters the requirements of the study design are described. In this chapter the different SA measurement methods will be presented. Since a mixed method design is preferred, different types of methods are selected and described with their pros and cons. Some measurement methods are described in detail. From these methods the most important requirements for SA measurement methods are evaluated: validity, sensitivity, reliability and intrusiveness. Finally the requirements for the measurement environment are given.

![Figure 13: This chapter contributes to the experimental set-up](image)

5.1 Overview

As mentioned in the introduction, the measurement methods can be divided in several categories. In Table 3 an overview of the SA measurement methods are given. In the left column the categorization of methods is given according to Salmon (2006) these methods are explained in chapters 5.1.1 until 5.1.7. The categorization is compared to the categorization of Endsley. Endsley categorized the methods in four different categories: process indices, direct measures, behavioural measures and performance measures (Endsley et al., 2000a). Process indices examine how people process different elements in the environment (Matthews et al., 2000). A direct measure of SA taps into the perceptions of the participant, an indirect measure infers the operators’ SA (Endsley, 1995a). Behavioural measures use observers to determine the SA based on the behaviour of the participant. Performance measure tries to determine SA based on the performance of the participant (Matthews et al., 2000).
Table 3: Overview of SA measurement methods

<table>
<thead>
<tr>
<th></th>
<th>Process indices</th>
<th>Direct/indirect</th>
<th>Behavioural measure</th>
<th>Performance measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA requirement analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Freeze probe techniques</td>
<td></td>
<td>Direct</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Real-time probe techniques</td>
<td></td>
<td>Direct</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-rating techniques</td>
<td></td>
<td>Direct</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observer-rating techniques</td>
<td></td>
<td>Indirect</td>
<td>Behavioural measure</td>
<td></td>
</tr>
<tr>
<td>Performance measures</td>
<td>Indirect</td>
<td></td>
<td>Performance measure</td>
<td></td>
</tr>
<tr>
<td>Process indices</td>
<td>Process indices</td>
<td>Indirect</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5.1.1 SA requirement analysis

Usually, subject matter experts (SME’s) are interviewed in order to develop a SA requirement analysis. This analysis is used as an input for the SA assessment technique (Salmon et al., 2006). It produces a list of the SA requirements that an operator should know.

5.1.2 Freeze probe techniques

Freeze probe techniques use freezes during simulation to ask questions to the operator. Queries are asked during a freeze about elements of the situation. Displays and the cockpit are usually withdrawn from the operator’s sight. The operator will answer the questions based on his or her memory (Salmon et al., 2006). The questions are not asked during the task itself therefore the operator cannot focus his awareness to the elements mentioned in the questions. After answering the questions the simulator resumes the simulation from the point where it was frozen. The advantage of these methods is the direct, fact-based approach. No opinions of SME’s are used in this method. A disadvantage of this method is its intrusiveness, since the simulator is frozen it interrupts the tasks of the operator. The operator’s SA is minimally influenced when different questions on each freeze are asked and freeze moments are timed as randomly as possible. This way prevents the operator preparing answers to the questions.

5.1.3 Real-time probe techniques

Real-time probe techniques are similar to the freeze probe techniques. The difference is in the moment the questions are asked. In real-time probe techniques the questions are asked during the task performance, the task is not paused. This has a lower level of intrusiveness than the freezing technique. The answer time and content are used to measure the SA of the driver (Salmon et al., 2006). This method is very usable when a task is not simulated or when it cannot be frozen (Endsley, 2000b). Although this method is not as intrusive as the freeze method, it still has a minor intrusiveness since the operator has to answer questions during the execution of its tasks. Also, when questions are asked during the execution of the tasks, the attention of the operator is directed at certain elements (Salmon et al., 2006).
5.1.4 Self-rating techniques
Obviously, in self-rating techniques the operator has to rate his own SA. This can be done via answering questions post-trial. This method is easy and cheap to use, and is not intrusive since it is post-trial. Disadvantages are that people might over- or underestimate their own SA and people tend to link their SA with their performance (Salmon et al., 2009, Endsley et al., 2003).

5.1.5 Observer-rating techniques
In observer-rating techniques, SME’s give a rating to the performance of the participant and their SA. This method is often used in the field. It is a non-intrusive method, but it is very difficult to rate a SA of somebody else. Often multiple SME’s are needed to give a balanced rating (Salmon et al., 2009).

5.1.6 Performance measures
Another non-intrusive method is performance measures. This can be measured during the execution of the task (Salmon et al., 2009). Examples of these measures are: amount of kills for a soldier, or amount of times that a train driver passes a red signal. The problem with this technique is that performance and SA is not always linked to each other; somebody might have a safe performance despite of a bad SA and vice versa (Salmon et al., 2009). Since it is an in-direct SA measurement method, it is not validated as an SA assessment method. Often, performance is measured during a test anyway, therefore it can be used as a reference for other measurement methods (Salmon et al., 2006).

5.1.7 Process indices
Process indices measure processes that participants use for developing SA (Salmon et al., 2009). The movement of the eye is an example of a process index. Some of these methods might be intrusive, for example when wearing an eye tracker some people feel limited in their movements of body. Problems with these techniques is the difficulty to relate it to the overall SA.

One method of each technique will be described in the next subchapters. The selection for each method was based on the validation, reliability, sensitivity and intrusiveness. The method should be able to measure SA in a simulator. An extensive overview of the different measurement methods can be found in (Salmon et al., 2006) and (Lo and Meijer, 2013)

5.2 GDTA

Description
A goal directed task analysis (GDTA) can be used as an SA-requirement analysis. It is developed to sort out which information an operator needs in order to make decisions when executing its job. The GDTA is developed by means of interviewing and observing operators and SME’s using instructions and manuals (Endsley et al., 2003). In the GDTA the overall goal is given, which is then divided in major goals and sub goals. These sub goals lead to decisions that should be made by an operator, from these decisions the SA requirements can be developed (Endsley et al., 2003).

The researcher should make one GDTA from all interviews, by categorising the answers. First only the goals should be given; these are items that require cognitive effort of the operator (Endsley et al., 2003). In order to achieve these goals an operator should make decisions. These decisions are given
beneath the goals in the hierarchy. Finally, the SA requirements are developed by means of identifying all the information that is needed to make the decision (Endsley et al., 2003).

The GD TA should be validated when it is developed. Validation can be done by asking expert for missing items, or by using verbal protocol in a simulation. In a verbal protocol the operator is asked to think aloud while he is operating (Endsley et al., 2003).

5.3 SAGAT

Description
The Situation Awareness Global Assessment Technique (SAGAT) is the best-known freeze probe technique. It is a widely used method to measure SA in human-in-the-loop simulators. In this method a simulation is frozen at randomly selected times in which the participant should answer detailed questions about the situation (Endsley, 2000a). It is an objective, direct method, since fact-based questions will be asked without any interpretation of an observer. Questions can be derived from the SA requirements as can be written in the Goal Directed Task Analysis (GDTA) (Endsley, 2000a) or Hierarchical Task Analysis (HTA) (Stanton, 2006). Although not required, a computer can be used to ask the queries. The participant can answer the question by clicking the answers or pointing at objects on maps, this minimizes the answering and analysing time. Also when a computer is used, only queries can be used which are appropriate for that moment (Endsley, 2000a). Query selection should be based on the capabilities of the simulator, the scenarios and the objective of the test (Endsley, 2000a). A broad range of queries should be asked in order to be sensitive enough to find changes in SA (Vidulich, 2000). Also, narrowly asked questions might shift the attention of the participants to only the elements of which questions are asked. Therefore queries about the important elements for the scenario should be combined with random queries about less important elements (Endsley, 2000a).

The freezes should be randomly timed, so that the participant cannot prepare for a freeze. A maximum of 5 minutes per freeze is recommended to make sure no memory loss will occur. The time between two freezes should be a minimal of 1 minute. There is no effect measured about the amount of times a simulation is frozen(Endsley, 2000a).

Training of participants is recommended, to teach them how to answer the questions. This is done in order to eliminate wrong answers because they did not understand the procedures or questions (Endsley, 2000a). Participants should always try to answer all the questions, even if they do not know the answer. By doing this, the participant also has to use his schemata when answering questions. This gives a complete picture of his SA (Endsley, 2000a).

The answers are rated as good (correct) or wrong. An interval of correct answers should be made when an absolute number is asked. A statistical test can be used to look at statistical differences between scenarios (Endsley, 2000a). Ideally, each query should be separately evaluated. Each query should be asked 30 to 60 times when a within-subjects design is used (Endsley, 2000a).
Validity
A lot of studies have proven the validity of SAGAT (Jones and Kaber, 2004, Endsley, 2000a).

Sensitivity
Different studies have proven that SAGAT is sensitive enough to measure differences between designs (Vidulich, 2000, Endsley et al., 2000c, Strater et al., 2001). Vidulich (2000) found that SAGAT is sensitive when the asked questions are broad enough; that is, not narrowed down to specific SA elements.

Reliability
Various researches have been executed about the reliability of SAGAT. Most studies proved that SAGAT is a reliable method (Endsley and Bolstad, 1994, Collier and Folleso, 1995, Gugerty, 1997). Only Fracker (1991) concluded that SAGAT has a low reliability, but Endsley (2000a) pointed out that this might be the effect of the used absolute error, the low-fidelity simulator and the inexperienced participants.

Intrusiveness
Also the intrusiveness of SAGAT has extensively been researched. Although some researchers still doubt that SAGAT is intrusive (Sarter and Woods, 1991), a lot of studies have concluded that there is no relation between the performance of the participants and the use of SAGAT (Endsley, 1990, Endsley, 1995a, Endsley, 1988b).

5.4 SPAM

Description
The Situation Present Assessment Method (SPAM) is a real-time probe technique; that is, during a situation questions are asked about that situation. Memory problems that occur with offline methods do not occur in this method. The situation is still present and the operator keeps performing his task (Durso and Dattel, 2004).

SME’s make the queries that have to be answered. When SME’s are also used during the test, they can select which question is applicable to the situation (Durso and Dattel, 2004).

The response time and accuracy of the answer is taken as an indicator of the SA. Operators who answer quick and correct are assumed to know the right answer and therefore have a developed SA. It is assumed that operators who answer slowly but correct did not know the answer and had to look at it (Durso and Dattel, 2004). Durso and Dattel give three advantages of SPAM: First, response time gives a statistically sensitive measure. Second, response time also allows to measure successful SA, instead of only unsuccessful SA. Third, understanding of the situation is best when the situation is present. Fourth, the method can be used in the field.

Every time before a question is asked, the researcher asks if the operator is ready to answer the question. This is to make sure that the response time is not correlated with the workload; when the workload is high response time will be longer (Durso and Dattel, 2004).

Validity
Several studies stated that SPAM is a valid method (Durso and Dattel, 2004, Durso et al., 1999).
**Sensitivity**
SPAM was capable to sense differences between different scenario’s that was correlated with the performance (Durso and Dattel, 2004).

**Reliability**
Several studies used SPAM as a method for assessing SA. All studies found a correlation between the performance (Durso and Dattel, 2004).

**Intrusiveness**
SPAM is an intrusive method since the questions are asked during the task performance, also the attention might be directed to specific SA elements (Salmon et al., 2006). In the research from Durso and Dattel (2004) performance of SPAM measurements were compared with the performance without measurement, no differences were found.

### 5.5 MARS

**Description**
Although SART is the most known self-rating method; some studies suggested that it lacks construct validity. That is, in this study the SART outcome does not correlate with the SAGAT outcome and the performance (Salmon et al., 2009). Therefore another self-rating method is chosen; Mission Awareness Rating Scale (MARS).

MARS was developed for the measuring the SA of infantry missions. It is based on Crew Awareness Rating Scale (CARS), in which participants answer two sets of four questions. The first three questions are related to Endsley’s three SA levels, the last question is about how well the participant was able to identify his goals for his current situation. The second set of questions is based on the workload of the situation (Matthews and Beal, 2002). MARS is a quick and easy method, which does not require a lot of training, and is less intrusive than online techniques (Salmon et al., 2006). Since MARS is a self-rating technique, it is also subject to the same problems: the outcome can be correlated with the performance and people tend to over- or underestimate their SA (Salmon et al., 2009).

**Validity**
Matthews et al. (2002) also concluded that MARS was a valid method, although they also suggested that more research is needed to validate MARS. He suggested doing additional research in which the MARS outcome can be compared with SAGAT.

**Sensitivity**
Mars was capable to measure differences in SA in a SA research about different night vision goggles. In this same research the authors concluded that high levels of SA were more difficult to rate than lower level SA (Matthews and Beal, 2002).

**Reliability**
Only two researches are executed with MARS. Therefore nothing can be said about its reliability.
Intrusiveness
MARS is a post-trial self-assessment method. The participant is not disturbed during the simulation or mission. Therefore the method is not intrusive.

5.6 SABARS

Description
Situation Awareness Behaviorally Anchored Rating Scale (SABARS) is an observer rated SA measurement method; it uses independent observers for rating the behaviours of the participants. It was designed to assess the SA of infantry soldiers and their leaders in the field (Matthews and Beal, 2002). The method does not rate actual SA, but behaviours that indicate SA (Strater et al., 2001). A selection of the SA requirement analysis should be made to construct the SABARS questions. On a five-point scale the observer can rate the behaviour of the participant (Strater et al., 2001). In previous studies SABARS was evaluated as an easy to use method for observers (Matthews and Beal, 2002).

Salmon et al. (2006) summarized the advantages and disadvantages: the biggest advantage of the method is that it is not intrusive. The three disadvantages are: some researchers doubt about the capability of observers to rate the SA of a participant in detail, nobody can look into somebody’s head. When participants are observed, it might influence the behaviour of the participants and therefore influence the measurement. Finally, a lot of SME’s and access to field settings are required for a valid measurement.

Validity
Two studies confirmed that SABARS is a valid study: Strater et al. (2001) concluded that several SABARS factors were related to the SAGAT outcome and experience levels of the platoon leaders, this study was executed in a simulator. Matthews et al. (2002) also concluded that SABARS was a valid method, although they also suggested that more research is needed to validate SABARS. This research was executed in the field.

Sensitivity
SABARS was sensitive for the experience level of the platoon leaders, just like the SAGAT method in that research (Matthews and Beal, 2002). In the research of Strater et al. (2001) nothing is explicitly said about the sensitivity, but since the outcome of SABARS was related to the SAGAT outcomes it can be concluded that the method was sensitive.

Reliability
Matthews et al. (2002) stated that more research is needed to establish the reliability. In the mentioned research only one observer was used, therefore nothing can be said about its reliability (Strater et al., 2001).

Intrusiveness
This method is not intrusive for the participants (Salmon et al., 2006).
5.7 TT-STS

Description
Useable process indices for simulator studies did not exist for railway – train situations. The amount of SPAD’s is often used as a performance indicator, but since the probability of a SPAD occurrence during a simulator study is very low, a new performance measure is used. As a performance measure a new indicator has been developed especially for the railway sector: TT-STS (Dutch: tijd tot passage stoptonend sein, English: TT-SPAD, time to signal passed at danger). This is the remaining time until it is irrevocable to pass a red signal.

In Figure 14 the TT-STS is schematically drawn. In this figure a train drives from left to right. When the TT-STS is 0 the train driver brakes maximal, since he has to stop before the red signal. The TT-STS would be negative if the train driver started braking later, then it is impossible to stop before the red signal. TT-STS is positive on the left side of the zero line; if he starts braking maximal before this line he will stop (far) before the red signal. This measure is derived from the time-to-collision (TTC) indicator used for traffic. TTC is the time left before two vehicles collide when nobody takes action (Hayward, 1972). “TTC has proven to be an effective measure for rating the severity of traffic conflicts and for discriminating critical from normal behaviour” (Van der Horst and Hogema, 1994).

TT-STS is calculated by the formula: (distance to red signal – minimum braking distance)/ speed (Van Luipen and Schotsman, 2013). TT-STS is calculated continuously and therefore has infinite measurement points. The minimal TT-STS data are compared between designs; the data is plotted in distributions, the distributions are then compared with each other.

Validity
TT-STS is a performance measure and therefore an in-direct method of SA. Therefore it does not measure SA. Performance is often correlated with SA.

Sensitivity
The sensitivity of the system is dependent on the equipment and software used. Most equipment will suffice.
Reliability
The reliability of the system is based on the equipment and software used. Most equipment will suffice.

Intrusiveness
TT-STS is measured simultaneously when the train driver is driving a simulator, without interrupting the train driver. Performance indices are a non-intrusive method (Salmon et al., 2006).

5.8 Eye tracker

Description
The eye tracker is a widely used method for measuring SA as a process instead of a product (Salmon et al., 2006). An often-heard comment on eye-tracker as an indicator for SA is the “look but failed to see” phenomenon as described by Brown (2001). Van de Merwe et al. (2012) are aware of this phenomenon but are reasonably confident that when an eye is focussed for a specific period on a location, the attention is located to that location. They reference to Shinar (2008) for a discussion.

Eye-trackers are used in previous SA measurements. The percentage of time that a person fixates at an area of interest was found as an important measure for SA, and the number of fixations was an indicator for the amount of errors made (Moore and Gugerty, 2010). In another research the fixation rate and the dwell time was found as a useful indicator for SA (van de Merwe et al., 2012).

There are some disadvantages of using an eye tracker (Salmon et al., 2006): The equipment is difficult to operate, the analysis is very time consuming, and it can only measure SA indirect.

Validity
No validation studies are found of the eye-tracker. Eye-tracker measurements are an in-direct measure (Salmon et al., 2006), and therefore does not measure SA. In the research of Van de Merwe et al. (2012) the outcome of the eye-tracker measurements were compared with the CARS outcome. The output of the methods correlated with each other.

Sensitivity
Several studies of eye-tracker measurements indicated that it was capable of sensing differences between different scenarios (e.g. (van de Merwe et al., 2012, Moore and Gugerty, 2010, Ratwani et al., 2010)).

Reliability
Several studies of eye-tracker measurements indicated that the eye-tracker was a useful indicator for SA (e.g. (van de Merwe et al., 2012, Moore and Gugerty, 2010, Ratwani et al., 2010)).

Intrusiveness
The eye-tracker is a non-intrusive method (van de Merwe et al., 2012).
5.9 Measurement environment

SA measurements can be executed in different simulators. Low-fidelity simulators can be used at an early stage when new designs are developed to see initial problems. At a later stage more detailed simulators are wanted for their broader range of environmental factors. The effectiveness of evaluating designs is strongly influenced by the quality and realism of the simulations (Endsley et al., 2003). Therefore, the simulator should be as real as possible. However, high-fidelity simulators are not always available. The simulator can then be tested if it is applicable for measuring SA.

An important aspect of a simulator is controllability of scenarios. A simulator cannot achieve the realism of a field-test, but it can be controlled (Endsley et al., 2003). The field-test is a useable test environment when the control of the situation is not very important for the measurements.

Some measurement methods have specific requirements of the test environment. When using SAGAT, a simulator has to be able to freeze and un-freeze, a field-test is therefore not applicable. Since SAGAT can ask detailed information about the situation, a high-fidelity simulator is recommended. Previous SAGAT measurements used high- and medium-fidelity simulators (Endsley, 1995a). When these simulators are not available, it is recommendable to choose another measurement.

5.10 Conclusion

In this chapter different SA measurement methods have been described. The problem with the measurement methods is that they are not all validated as SA measurement methods. Since it is recommended to have a mixed method design, multiple methods can be used. SAGAT is the best-known method for objective measuring SA. It is validated, sensitive, reliable, and despite of the commentary of its intrusiveness no relation is found between performance and the use of SAGAT. This method is recommended as best to use.

To make the outcome of the measurements more reliable other methods can be combined with SAGAT; if the outcomes of all methods give the same results, it advances the reliability. SAGAT can be combined with: a post-trial subjective measurement method such as MARS, an observer rating method such as SABARS, performance measures such as TT-STS, or process indices such as an eye-tracker. All these methods can be used at the same test run. A real-time probe technique can also be used, but it is recommendable to do this in separate test runs since these methods are intrusive. At least one method should be combined with SAGAT. The MARS is the easiest to combine with SAGAT, and the methods do not influence each other. Therefore, MARS is recommended to use in combination with SAGAT.

The test environment is best when it is as realistic as possible and controllable. When it is not possible to get a high-fidelity simulator or a field-test, the simulator should be tested for measuring SA.
In the previous chapters the requirements of measuring SA for train drivers are determined. In this chapter executed SA tests are described. This case study is performed to check if SA measurements are useable to take safety decisions. The SA of train drivers is measured in the DSSU project (DoorstroomStation Utrecht) using the objective SAGAT and the subjective MARS method in the NSO (Nederlandse Spoorwegen Opleidingen, NS Dutch railways education) simulator in Amersfoort. The purpose of these measurements was to identify differences in train drivers’ SA for different infrastructure layouts. On top of this research, the SA was also measured in the low-fidelity simulator MATRICS, to see if the same trends in SA can be detected.

![Figure 15: This chapter explains the executed SA tests](image)

### 6.1 DSSU description

The need for train capacity is growing (OVPro.nl, 2012), ProRail is expanding the capacity in several ways. One of them is the DSSU-project. This project is aiming to improve the capacity, punctuality, flow, and robustness at disturbances in the Utrecht area. To achieve this solutions are designed; corridors are unbundled, an extra platform is built, and switches are used to pass with a higher speed. The unbundling can be seen in Figure 16 and Figure 17. Figure 16 shows the old infrastructure layout with a lot of switches and crossing tracks. Figure 17 shows the new DSSU layout, in which a lot of switches are removed, corridors are added and thereby a lot of crossing tracks disappeared. A new signalling design for the sections close to the train station is made to allow trains to enter the station faster following the previous trains. Therefore shorter signal blocks are made and DDDR (Dynamisch Dubbele Doorgaande Remming, freely translated: dynamic double non-interrupted braking) is applied (ProRail, 2013b).
In the old signalling design the length of the blocks are set depending on the distance needed to brake from the speed limit to zero. In the old situation the following signs can occur:

- Green, speed limit at sign is the same as the section speed limit
- Green blinking, speed limit at sign is 40 km/h
- Green + blinking number, speed limit at sign is blinking number
- Yellow, at next sign the speed limit is 40 km/h
- Yellow + number, at next sign the speed limit is the number on the sign
- Yellow + blinking number, the speed limit of the blinking number cannot be reached at the next sign. Braking should not be interrupted if next sign allows a lower speed than your speed. This situation does not occur very often.
- Red, stop

In DDRD the yellow + blinking number sign can occur in succession of each other, without the condition that the numbers are decreasing. The signal is used for regulating speed for safety purposes and creating a flow. In Figure 18 DDRD is schematically drawn. The upper sign succession
gives a regular non-interrupted braking succession. The middle sign succession gives a double non-interrupted braking succession. The bottom sign succession gives a dynamic double non-interrupted (DDDR) braking succession. This DDDR might have negative consequences for the behaviour, performance, alertness and fatigue of the train driver. Especially the expectations of the train drivers of the next sign can change (Zon et al., 2013). Because of these reasons, the management of ProRail and NS wanted to investigate if the train driver will behave differently in the new situation. If they behave differently they also wanted to know the reasons of this behaviour. SA was measured to see if the SA in the DSSU project would be different than it is in the current situation. A complete SA is needed for the train driver to be able to drive the train safely.

![Different signal successions](image)

**Figure 18: Different signal successions (Zon et al., 2013)**

### 6.2 Simulator

Eight train drivers of the NS drove four different scenarios in the full scope simulators of the NSO. These simulators are used by the NSO for initial and revision training of rail drivers. Two simulators are available; the SLT and the VIRM, these are two types of trains used by the NS. The cabin of the simulators is an exact copy of a real train. In Figure 19 the interior is showed of the SLT simulator. The simulator also has motion, to simulate accelerations, braking, corners and bumps in the rail track. The layout of the infrastructure is an exact copy of the reality; e.g. placing of the signs, speeds and routes of the rail. Buildings near train stations are modelled, the train driver can orientate on these buildings as part of his route knowledge. Furthermore the train characteristics are modelled. The observer is placed in the room next to the simulator; the observer is an educator who also acts as the train dispatcher. In this simulation environment the train driver is in the most realistic situation possible.
6.3 Scenarios

All the train drivers drove four scenarios. These were a combination of signalling designs and nuisance:

- S1: The old signal design without nuisance
- S2: The old signal design with nuisance
- S3: The DSSU design without nuisance
- S4: The DSSU design with nuisance

The scenarios were chosen in collaboration with the DSSU project management to make sure they would accept these scenarios as representative for DSSU. The nuisance was based on representative situations; the nuisance was caused by a delayed freight train.

The scenarios were offered in a different order to each driver to prevent a similar learning curve. Each driver got a short briefing explaining the purpose of the measurements and the difference in infrastructure (shorter block and different placing off the signs), nothing was told about the nuisance or the moment the measure was in action. They did not know which scenario they were driving or which signalling design they would get.

During the SA measurements four train drivers drove the SLT, the other four the VIRM simulator. At the same day another test was also taken. In this test workload data was collected of train drivers via an eye-tracker and hart rate measurement. The scenarios of this test were almost identical to the scenarios used in the SA measurements. Each day two train drivers were active. The first train driver started with the SA measurements, the second with the workload measurements. When the drivers completed the four scenarios there was a break. After the break the drivers switched places. Therefore the driver knew the scenarios after the dining break because he had already driven the scenarios before the break, but the sequence of the scenarios was different.

In every scenario during the SA-measurement a strange event occurred. The events were: a pink rabbit next to the rail, a broken sign at the adjacent rail, a GSM-R (GSM-Railway) failure, and an ATB-error (Automatische trein beïnvloeding; automatic train stopping device). The ATB system
automatically stops a train if a train driver drives to quickly. These four events were the only difference in scenarios between the SA measurements and the workload measurements.

6.4 Participants

The participants of the simulations were all male train drivers of the NS, aged between 26 and 63 years. They all have route knowledge on the simulated route. The participants were selected based on availability and route knowledge. Route knowledge is mandatory to drive a train on Dutch rail.

6.5 SA measurement set up

The SAGAT method and a modified MARS method were used to measure the SA of the train drivers. Although the eye-tracker could also be used for measuring SA, this was not done because of the very long processing time. For the SAGAT method, questions were derived from the GDTA (Goal directed task analysis).

The GDTA was developed by combining Dutch and British documents of train driver’s occupational competences, which was converted into tasks. After that, a Dutch train driver and an instructor separately validated the GDTA. From this GDTA all elements that had a direct relation with signalling and speed were identified and each element was converted into a total of fifteen questions. Twenty-six extra questions were made up, which the train driver should be able to answer, but had no relation with the signalling or speed. These questions are the ‘distraction’ questions, to make sure the drivers did not only focus on the signalling and speed questions.

All the questions were multiple-choice in order to let the train driver answer all the questions as quickly as possible. A train driver (not one of the test drivers) checked the questions if they were understandable, using the right words, and not multi-interpretable, before the questions were finalized. The questionnaire is given in Appendix D: SAGAT question forms DSSU.

During every scenario the simulator was stopped twice. First a towel was thrown over the train driver’s head to blind his sight and immediately the simulation was paused. The train driver was asked to turn his chair around and answer the questionnaire on paper. When the questions were completed, the scenario was resumed at the same place and with the same situation as before the interruption. At the same time, the observer also answered the questions, while he could see the drivers’ simulation screen to check the answers. The answers of the observer where used to check the answers of the train drivers.

The purpose of the research was to check whether there are any differences in SA of the train drivers between the old signalling and DSSU signalling. Therefore the questions in the old signal design without nuisance where the same as in the DSSU design without nuisance. The questions in the old signal design with nuisance where the same as in the DSSU design with nuisance. By doing this, the best comparison could be made between the two signalling designs.
6.6 Data processing

All the scenarios were taped on video to check the answers of the observers. Then the corrected answers of the observers were compared with the answers of the train drivers. Most answers were right or wrong, but some other answers are doubtful. The questions about speed were answered on a scale. This scale was between 0 and 130 km/h, an X should be set on the speed they thought was correct. The advantage of this method is that participants could quickly answer the question and were not forced to answer in categories. For example, when the train driver was driving 81 km/h and the possible categories were 60-80 km/h and 81-90 km/h only the second answer was right. But if the train driver thought he was driving 80 km/h, he was only 1 km/h next to the actual speed, is the first answer wrong? To solve this problem the speed should be set on a scale. The disadvantage of this system is that the given answers cannot be determined precisely, here for the given answers were rated right if it was in a range of 9 km/h of the actual speed. This range was chosen based on common sense. Most people do not know their actual driving speed in a car either: when the speed limit is 120 km/h you can easily drive 110 km/h or 130 km/h when not constantly looking at your speedometer. Driving 105 km/h or 135 km/h is not very common; normally you already checked your speedometer before you drive that speed. Train drivers agreed that this situation is the same as on the rail. For one question, multiple answers were right. In this question was asked for the types of breaks under the train. When one of the answers was not filled in, the question was rated wrong. Since the train driver should be able to correctly answer the ‘distraction’ questions, both the questions of the signalling and the speed as the ‘distraction’ questions were used in the analysis.

6.7 Data analysis

A comparison is made between the amounts of correct answers in the two infrastructures. First a comparison is made between the SAGAT answers and the signalling design, independent of the nuisance. Secondly a comparison is made between the SAGAT answers and the nuisance, independent of the signalling design. Thirdly a comparison is made between the SAGAT answers and the signalling strategy per nuisance scenario.

In Figure 20 the total amount of correct (good) answers in the two signalling designs are compared with each other from all train drivers. In the upper figure the amount of wrong answers is given, in the bottom figure the amount of correct (good) answers is given. No distinction is made in nuisance level in this figure. The blue bar represents scenario S1 + S2, the green bar represents scenario S3 + S4. From the figure no difference in amount of correct (good) answers between the two infra conditions is visible. Figure 21 is given to see all the individual scores off the drivers, to make sure no individual train driver behaved very different.
As can be seen in Figure 21 no extreme values are visible from any train driver; none of the train drivers performed a lot better in either S1 + S2 or S3 + S4. A chi-square test was also executed. A comparison was made between the signalling design and the answers on the SAGAT questions. The outcome of the chi-square test was that there was no statistical relation between the signalling design and the SAGAT answers.

Since no difference in SA was found between the different signalling designs, it might be possible that the measurement method is not sensitive enough. Therefore the differences in SA between different nuisance levels are also tested. This is given in Figure 22, the left bar represents scenarios S1+ S3, the right bar represents S2 + S4. It seems that there is a big difference in SAGAT answers between the nuisance scenarios. The chi-square test also shows that there is a statistical relation between the SAGAT answers and the nuisance. A post-hoc test gave a phi of 0.167, which means that the relation is very weak.
Since there is a relation found between the SAGAT answers and nuisance, it is very interesting to see if the relation is stronger in the old signalling strategy or the new signalling strategy. In Figure 23 are all scenarios given in sequence in separate bars; S1 is the left bar, S4 is the right bar. The figure shows that the difference in SAGAT answers with nuisance is bigger in the new signalling strategy. A chi-square test was executed for both signalling designs, both tests showed that there is a statistical relation between the SAGAT answers and the nuisance. The post-hoc test gave a phi of 0.148 for the old signalling design and a phi of 0.184 for the DSSU signalling design. So both signalling designs had a very weak relation.

Since the SAGAT questions are kept the same between two signalling designs and different between two nuisance scenarios, it should also be checked if the wrong answered questions of the nuisance scenarios are answered correctly in the without nuisance scenarios. If this is the case the statistical relation could be explained by the use of different questions.
<table>
<thead>
<tr>
<th>Question ID</th>
<th>2a2</th>
<th>2a3</th>
<th>2a8</th>
<th>2a11</th>
<th>2b4</th>
<th>2b9</th>
<th>2a7</th>
<th>2a9</th>
</tr>
</thead>
<tbody>
<tr>
<td>With nuisance</td>
<td>3</td>
<td>5</td>
<td>7</td>
<td>6</td>
<td>9</td>
<td>11</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>Without nuisance</td>
<td>6</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

In Table 4 an overview is given to show the questions that are answered wrong five times or more in total. The question ID is related to the question form and the question; e.g. 2a3 is question form 2a question 3. These forms are shown in Appendix D: SAGAT question forms DSSU. As can be seen in Table 4 most questions are answered more often wrong in the scenarios with nuisance. Only question 2a2 is answered wrong more often in the scenario without nuisance. Question 2a7 and 2a9 were not asked in the scenarios without nuisance, so no comparison is possible for these questions.

Two questions stand out because of the high amount of questions answered wrong: 2b4 and 2b9. Question 2b4 is: Is the train delayed by more than three minutes? Questions 2b9 is: What do you expect that the coming signs will allow you? With the answers: I can speed up / I can keep this speed / I will have to break. Both questions have two properties. First, both questions were asked in the first stop in the scenario without nuisance and in the second stop with nuisance. Both questions are easier to answer in the first stop. In the first stop the delay was not very big yet, in the second stop it was. In the first stop it was very clear if there was a freight train in front of the train, therefore it is easy to answer if they could speed up or not. Second, both questions could not be checked on the video, because it was not visible. When checking the questions of the observers on the video, a lot of answers had to be corrected.

### 6.8 MARS

A mixed method design for measuring SA is used to make the answer more robust. The second method for measuring SA is the MARS method. This subjective method is not intrusive since the simulator does not have to be paused and questions are asked after the scenario ended. This method is very easy to answer for train drivers, but every train driver might answer the questions differently because some train drivers are more critical than others. Six questions were asked after every ride.

1. To what extend could you focus on: signs, the panel, traffic, during the ride?
2. Did you know everything you needed to know during the ride in order to drive the train in a good way?
3. I had a good understanding of what was happening.
4. I could predict well what would happen during the ride.
5. I could drive the train well (safely, punctual, according to the NS-handbook, energy efficient, comfortable for the passengers) with all the available information.
6. I am satisfied with the decisions I made.

The questions are related to the SA-levels as Endsley described; question one is about the perception of relevant information, question two and three are about the understanding and the meaning of that information, and question four, five and six is about the prediction of future events based on...
this information. All these questions were answered on a seven-point scale ranging from “not at all” (one) to “excellent” (seven). In Table 5 the outcome of the MARS measurement is given, all the values in the table are mean values. The train driver and the observer answered every question. This method is normally not used for observers; the SABARS method is more suitable for observer ratings. SABARS was also considered as a measurement method, but due to time constraints for developing SABARS questions this is not used at this occasion. By letting the observers answer the MARS questions is examined if MARS could also be used by observers to score the SA of operators.

In Figure 24 till Figure 29, all the outcomes of the MARS questions are given. All the figures are structured in the same manner: the four bars on the left represent the answers given by the train drivers, the bars on the right represent the answers given by the observers, the bars are in sequence of scenario numbers. In the figures are the error bars given in black; this represents the uncertainty of the measurements. Statistical tests were performed but all the outcomes were insignificant; e.g. there were no significant differences found between MARS scores dependent on infrastructure and nuisance. This is partly because only eight drivers attended the test. There are some visible differences though.

<table>
<thead>
<tr>
<th>Question</th>
<th>Answered by</th>
<th>Old infra without nuisance</th>
<th>Old infra with nuisance</th>
<th>DSSU without nuisance</th>
<th>DSSU with nuisance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question 1</td>
<td>Train driver</td>
<td>6.13</td>
<td>5.5</td>
<td>5.88</td>
<td>5.75</td>
</tr>
<tr>
<td>Observer</td>
<td>5.86</td>
<td>5.88</td>
<td>6</td>
<td>6.14</td>
<td></td>
</tr>
<tr>
<td>Question 2</td>
<td>Train driver</td>
<td>6.5</td>
<td>6.38</td>
<td>6.5</td>
<td>6.63</td>
</tr>
<tr>
<td>Observer</td>
<td>6.14</td>
<td>6</td>
<td>6.43</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Question 3</td>
<td>Train driver</td>
<td>6.38</td>
<td>6</td>
<td>6.25</td>
<td>5.75</td>
</tr>
<tr>
<td>Observer</td>
<td>6.43</td>
<td>5.63</td>
<td>6.14</td>
<td>5.71</td>
<td></td>
</tr>
<tr>
<td>Question 4</td>
<td>Train driver</td>
<td>5.75</td>
<td>4.75</td>
<td>6</td>
<td>4.88</td>
</tr>
<tr>
<td>Observer</td>
<td>6.29</td>
<td>5.5</td>
<td>6.14</td>
<td>5.71</td>
<td></td>
</tr>
<tr>
<td>Question 5</td>
<td>Train driver</td>
<td>6.38</td>
<td>5.25</td>
<td>6.5</td>
<td>4.75</td>
</tr>
<tr>
<td>Observer</td>
<td>6</td>
<td>5.88</td>
<td>6</td>
<td>5.86</td>
<td></td>
</tr>
<tr>
<td>Question 6</td>
<td>Train driver</td>
<td>6.25</td>
<td>6.25</td>
<td>6.63</td>
<td>6.13</td>
</tr>
<tr>
<td>Observer</td>
<td>6.71</td>
<td>6</td>
<td>6.43</td>
<td>6.14</td>
<td></td>
</tr>
</tbody>
</table>
In Figure 24 the first question shown. The train drivers scored this question better with nuisance in DSSU than in the old infrastructure. But without nuisance the old infrastructure scored better. A possible explanation for this is that in DSSU more signs are placed on a short distance. Without nuisance the speed of the train is high, here for signals are passed very frequently and therefore it is harder to focus on the signals in DSSU than in the old infrastructure. Signals are passed less frequent in the scenarios without nuisance. In this situation it can be very helpful for a train driver if he passes more signals. The observer did not score differently between nuisance levels.

The answers of DSSU with nuisance of question two and three, in Figure 25 and Figure 26, seems to be contradicting with each other. In question two the train driver rated the question better in the DSSU with nuisance, than in DSSU without nuisance. In question three this is the other way around; i.e. the train driver had a better understanding in DSSU with nuisance to drive his train in a good way.
but also had a worse understanding of what was going on. A possible explanation for this answer is that the signalling in DSSU was clearer for the train driver since he got more signals, but he did not understand why these signals occurred. The observers rated the questions in the same manner as the train drivers.

![Figure 26: MARS question 3](image)

Figure 26: MARS question 3

Question four in Figure 27 does not show any unexpected answers. The train driver could predict what would happen better in the scenarios without nuisance than in the scenarios with nuisance. The observer rated the question in the same manner as the train driver.

![Figure 27: MARS question 4](image)

Figure 27: MARS question 4
In question five in Figure 29 the difference between the scenarios with and without nuisance are a lot bigger for the train drivers than for the observers. The observers were not able to accurately rate the train drivers. It might be that the train driver was able to drive the train very well with all the possible information, but they might feel that they were not able to do that. Unfortunately this question was very broad; the reason why train drivers rated this question differently for the different nuisance levels is unknown. It might be the punctuality since the trains always had big delays in the scenarios with nuisance, or the energy efficiency since they stated this in the interviews after the simulations.

Question six in Figure 29 does not show any unexpected answers. The observer rated this question in the same manner as the train driver.

Overall the observers did not give the same absolute rates to the questions as the drivers did, but most questions were rated in the same manner as the train drivers did; by an optical comparison the trends of the scores seems to be the same. The observers did score the questions with smaller differences between scenarios than the train drivers did; i.e. their sensitivity of rating is smaller. Therefore, observers need more measurements to prove any statistical differences.
6.9 SAGAT vs. MARS

Since MARS could not statistically be tested on optical comparison was made. By an optical comparison the same trends seems to be appearing in SAGAT and MARS. In both methods no difference in SA can be seen between infrastructure layouts, but there is a difference between the nuisances. Both methods are suitable for the measurements, but the MARS method did not give any significant differences due to the small amount of train drivers available. The SAGAT method is more precise and based on facts. By answering a lot of fact based questions. Since the SAGAT method collected much more questions during the measurement, it is the only method that gave significant differences between groups. If more participants were available, the MARS method could be suitable as well. The MARS method is useable as a second SA measuring method, to check the SAGAT outcome. The outcome of SAGAT is now more robust, since the outcome of MARS is the same as the SAGAT outcome.

6.10 Comments

A couple of things should be addressed to take into account for the outcomes:

Two different observers rated the drivers. One observer rated six drivers, the other two drivers. For this reason the drivers are not uniformly rated. This makes the ratings of the observers less reliable. Using different observers does not have to be a problem if there are a lot of different observers.

The observer added the events that occurred in the scenarios during the scenario run. Therefore two mistakes occurred. The first mistake was that in one scenario the pink rabbit was added on the track instead of next to the track. Since the rabbit was very big this was almost like hitting a person. Luckily the train driver did not responded by panicking but the event did have an effect on his SA. The second mistake was that one observer added a different ATB-error. This error was much more intrusive because the train reacted by alarming the train driver. This also had a big effect on the SA of the driver.

In this test a small number of participants were available. Therefore it is very hard to put statistically significance to the test. Differences in SA can be see more clearly if more drivers were participating.

Only one route of the DSSU project was taken for the SA research. First was proposed to investigate the changing SA of different signal placing strategies based on non-existing routes, in order to make the results route independent. Since the tender phase of the DSSU project already started, this project had a very high priority and should therefore be researched first. To be able to execute this research quickly, existing routes of the simulator were used. Therefore the route was taken from the DSSU project, which was available in the simulator. Since the research is based on only one route, the generalizability and transferability of the results should be investigated.

Only NS train drivers were used in this research. They all have route knowledge and drive often on this route. Route knowledge is the knowledge of a specific route about all kinds of characteristics such as were signals are, what the allowed speed is, were corners are etcetera. A train driver should have route knowledge before he is allowed to drive on this route. Besides NS drivers, also temporary
workforce drives on these tracks. They do not drive here very often which can make the outcome of this research different.

Long-term results are not available. On the long-term expectations of the drivers may change. If a yellow sign occurs very often in a row, drivers might change their behaviour when approaching a yellow signal. Letting a driver every week drive one of the DSSU scenarios could show long-term effects, by analysing trends. A monitor program can also help to analyse long-term effects, but the disadvantage is that this can only be implemented when DSSU is already executed. This makes it difficult to act on these problems.

The scenarios were not designed to test if there are events in which DSSU would give a very bad SA and the old infrastructure would have the same SA. For example when the SA of DSSU would be very low if the sun was near the horizon and the SA of the old infrastructure was the same. If these scenarios were also integrated in the test, the worst case of the DSSU would be visible. If the SA also did not change in this scenario, the robustness of the answer would even be better. ProRail tests in normal operating conditions, during a sunny day, train drivers are expected to adjust their driving behaviour when an unsafe event occurs.

### 6.11 MATRICS

Prior to the research about SA and workload measurements another research was executed to investigate the TT-STS in the different infrastructure layouts. This was a quantitative research in the low fidelity simulator MATRICS as showed in Figure 30. The safety research of DSSU is more robust by combining the outcome of this TT-STS research and the outcome of the SA and workload measurements: a mixed method design. The quantitative data is used to give power to the answers; the qualitative data is used to get an insight in what is going on in the train drivers. Since this TT-STS research was executed anyway, it was also used to investigate if SA measurements would be possible in the MATRICS simulator.

A low fidelity simulator is not suited for SA measurements, but an experiment with this simulator is much easier to set up. If the results can be used to give an indication of the SA of train drivers, this simulator can help to quickly investigate if big changes in SA appear. In this simulation forty train drivers drove twenty-four scenarios. These scenarios were based on four routes, three infrastructure designs, and two levels of nuisance. Only the route, infrastructures, and nuisance levels that are used in the full scope simulator measurements are used to compare the results. Some of the MARS questions were used for measuring SA in MATRICS.

Unfortunately, none of the questions have significant differences between infrastructure layouts or nuisances. Visible comparison of the scores also did not identify any differences. MATRICS might be helpful if big differences in SA were found in the full scope simulator, but further research is necessary to prove this.
6.12 Conclusion

SAGAT and MARS were used to measure the SA of the train drivers. The case study indicated that SA does not seem to change significantly between the current situation and the DSSU situation. The SA did change between levels of nuisance; therefore we know that the chosen methods were sensitive enough to measure differences. The study should be repeated for a check on the reliability of the measurements. Only SAGAT was able to identify significant differences in SA, therefore it is recommended to include SAGAT in future SA measurements. MARS was not able to identify significant differences in SA, but an optical comparison with the SAGAT outcome showed a correlation between these methods. MARS might be capable to show significant differences in SA when more train drivers would be available. MARS is now used to give more robustness to the answers of the SAGAT outcome.

Measuring in the low fidelity simulator MATRICS did not give any differences in SA at all. A new test in MATRICS can be done if big differences in SA in the full scope simulator are found. If this is the case, MATRICS can be used to give an indication if SA should be measured in the full scope simulator.
By combining the outcome of the TT-STS, SA, and workload measurements, a robust answer is given to the safety case of DSSU: The probability of a SPAD decreases in DSSU compared to the current situation (Van Luipen and Schotsman, 2013). In Figure 32 is indicated how the three measurement methods combined indicated the probability of a SPAD. The TT-STS identified the probability of a SPAD, the SA and workload measurements gave a better insight in the behaviour of a train driver. Combined gave this the overall probability of a SPAD.

The SA measurements helped to give an insight in the brain of the train driver: it indicates how aware he is of the situation, which was an indicator for being capable to make the right decisions for driving the train safely. A bad SA can indicate safety issues of a design, because SA is needed to make an informed decision. When a train driver has a correct SA, this does not imply that the design is safe. It only indicates that the train driver is capable to make the right decisions. SA as a single indicator for safety can therefore only be used when the SA is bad. If the SA is correct it can make the outcome of a quantitative TT-STS research more robust.

Figure 32: Probability of a SPAD
Decision-making based on measurements of behavioural safety

René van der Meij
7 Generalizability and Transferability

Since project management decided that DSSU was safe to implement based on the results of chapter 6, strategic management now wants to know if DDR can be implemented everywhere in the Netherlands; i.e. they want to apply a change on system-level. Therefore this chapter describes if new SA measurements are needed and if the SA measurements can be used for DDR. As can be seen in Figure 33, this chapter describes the goal and the problem definition.

![Figure 33: This chapter contributes to the goal and problem definition of DDR]

7.1 Determination of SA in other projects

The research in chapter 6 was based on the project DSSU; one specific route was researched on changes in SA. Some options are available for determining the SA of other projects, such as for DDR in the Netherlands. The methods are ordered on reliability: one is the most reliable method, three the least reliable. But the most reliable method is also the most expensive and has the longest run time. Based on the expected risk for changing SA, the best method should be chosen. The options are:

1. Measure SA of every single project
2. Generalizing the results
3. Determining that SA is not changing enough at all, stop measuring SA

Measuring SA of every single project is expected to be excessive. Small differences in SA are not interesting to know, since SA can only be used if significant differences between designs are found. Only projects that expect to have an influence on the SA and which are not comparable with previous SA measurements should be measured separately.
Using these results for other projects would be very desirable. In order to do this, the generalizability should be investigated. Research results are often limited to the research conditions. It is therefore recommended to take requirements for generalizability and transferability into account before designing a behavioural safety study. Generalization might be more difficult if this is not taken into account before the study was executed. In this chapter generalizability is extensively described. To begin with a generalization theory will be discussed which contains a generalization and a decision study. Apart from this mathematic method three different generalization theories can be distinguished: statistical generalization, analytical generalization and transferability (Polit and Beck, 2010). Statistical generalization is used for sample-to-population generalization, which is used in quantitative research. The requirement for a sample to population generalization is a representative sample, since this is already extensively described in literature and only focuses on quantitative research, this is not further elaborated. Analytical generalization is used to generalize the results to a broader theory. It can be used in both quantitative and qualitative research. Transferability is also described as “case-to-case” (Firestone, 1993) or “other-setting” (Seddon and Scheepers, 2006) generalization. It is used in qualitative research (Polit and Beck, 2010).

The third possibility is that SA is reputed to never changing so much that it will influence the safety. If other routes are also investigated and the outcome is the same (SA does not change very much), than SA should not be taken into account in the first place. This can only be concluded if more routes are investigated with all kinds of changed characteristics in it.

### 7.2 Generalization theory

Generalizability (G) theory is used to determine the dependability of an observation, or how accurately a specific, observed sample of behaviour reflects actual behaviour under the range of all possible facets of the measurement (Shavelson and Webb, 1991, Chafouleas et al., 2010). Facets are characteristics of the measurement such as the occasion, items involved, and observers. In a G-study measurements are based on a sample of the universe of admissible observations, these are all the possible combinations of the levels of the facets. G-theory could also be described as a random effects theory; random facets are created by randomly sampling conditions of a measurement procedure (Shavelson and Webb, 2005).

The main difference between classical test theory and G-theory is the use of error variance; classical test theory only examines one source for error variance, G-theory can examine multiple sources of error variance simultaneously (Brennan, 2001, Shavelson and Webb, 1991). The standard errors of variance of different facets are analysed, statistical theory is not used (Brennan, 2001). Ideally, in a G-study the variance of the measurement is only dependent on the participant and the occasion if the conditions of the measurement are perfect (Chafouleas et al., 2010).

In a decision (D) study a universe of generalization is defined, this is the set of conditions to which a decision maker wants to generalize. In the D-study the error of the measurement is minimized on specific items in order to be able to generalize the results. It uses information of a G-study to do this. A G-study is used to estimate the variability of as many possible facets as possible (Shavelson and Webb, 2005). Decisions usually will be based on the mean over the multiple observations rather than
on a single observation. Two kinds of decisions can be made: relative and absolute decisions. In relative decisions, individuals are compared with each other. Absolute decisions are independent of others’ performance (Shavelson and Webb, 2005).

In a D-study the researcher should hypnotize the change of the generalizability coefficient and the dependability index under different circumstances to determine the ideal conditions for the measurement. If this is known, the coefficients of the most dependable circumstances can be tested. Relative interpretations of measurement outcomes are based on the generalizability coefficients. The dependability index is used in absolute interpretations (Chafouleas et al., 2010).

The design choices in G and D-studies differ from each other. Crossed designs are used in G-studies to avoid confounding of effects. In a crossed design all participants are tested on all possible combinations of facets that influence the outcome of the results. In D-studies, nested designs can be used. In a nested design each participant is only tested on one combination of facets (Shavelson and Webb, 2005). A fixed facet can also be used in a G-study if the decision maker does not want to generalize over a facet or when it is unreasonable to generalize beyond the condition. When there are not a lot of conditions it can also be feasible to include all the conditions in the measurement. When the G-study has a bigger universe of admissible observations than the D-study wants to generalize on, the levels of a facet might be reduced or facets might be ignored (Shavelson and Webb, 2005). In order to be able to make a correct generalization, all participants and facets which are used in the measurements should be representative of the universe of generalization (Chafouleas et al., 2010).

G theory is very useful if the facets to which should be generalized are known. For example, if in the DSSU study only should be generalized over the distances between two signals, only a sample of these facets should be included in the study. Obviously, if more facets are important for the generalization, multiple facets should be sampled in the study. In this method it is very important that multiple random samples can be collected. Therefore, multiple scenarios, or scenarios with multiple areas in which different measurements can take place, should be included in the study.

### 7.3 Analytical generalization

An often heard criticism is that qualitative research is based on small sample sizes and therefore statistical generalization is impossible (Rodon and Sesé, 2008). Because of that reason analytical generalization should be used instead of statistical generalization. In analytical generalization the results are generalized to a broader theory (Whitehead and Yin, 2003).

For analytic generalization, conceptualizations of processes and human experiences should be made, both in detail as in abstraction (Polit and Beck, 2010). Evidence should be provided to support the theory and conceptualizations. For generalization to a theory, the theory should be used to make predictions and thereafter confirm those predictions. Scope conditions can be given to the theory that limits the conditions in which the theory may be applied (Firestone, 1993). In chapter 7.7 is explained how analytical generalization can be enhanced.
A theory should be developed before the analytical generalization can be used. For the DSSU project this can be: DSSU will not be less safe than the old situation because train drivers get more information and therefore will be more aware of the new situation, different length blocks do not have an impact on the safety. Alternatively a rival theory can be developed which is also feasible; DSSU will be unsafe because train drivers get an overload on information, short lengths of the blocks have a negative impact on the safety. By using two rival and feasible theories, the outcome of the study will be more potent (Whitehead and Yin, 2003).

### 7.4 Transferability

Another often heard criticism is that the results of qualitative research are context-bound (Rodon and Sesé, 2008). Therefore Seddon and Scheepers (2006) use “other-setting generalization”. This can be explained as “the researcher’s act of arguing, based on the representativeness of a sample, that there is a reasonable expectation that a knowledge claim already believed to be true in one or more settings is also true in other clearly defined settings” (Seddon and Scheepers, 2006, p. 1142). Firestone (1993) describes this process as case-to-case transfer. It is very closely related to analytical generalization. The difference between analytical generalization and case-to-case transfer could be described as follows: “Analytic generalization is facilitated by specifying the conditions under which a study is done and their relevance to multiple theories. That knowledge is used to create controls, quasi-experimental designs, or replications that strengthen generalization. Case-to-case transfer is enhanced by thick description that allows assessment of the applicability of study conclusions to one’s own situation” (Firestone, 1993, p.18).

The DSSU project can be generalized using case-to-case transfer. Only one route of the DSSU project was included in the SA research. Important settings of this research can be identified in order that the results can be used at other locations as well. The important elements are identified in chapter 7.8 and 7.9. The results may be applied to other locations that have the same elements, without any extra elements.

### 7.5 Relation between generalizability and precision of outcomes

The precision of the outcome of the safety research does not have to be very detailed, because the criterion of the DSSU project was “at least as safe” as the current situation. Van Luipen and Schotsman (2013) argued in the DSSU study that the generalizability of the answer is higher when the precision of the outcome is lower, this is schematically given in Figure 34.

Two researches were executed in the DSSU research. The first research examined the exposure of the red signs. The conclusion of this research is that the exposure of red signals and unplanned stops decreased with 70% (Van Luipen and Schotsman, 2013). The exposure of the yellow signals decreased with 70% in the scenarios without nuisance, and decreased with 38% in the scenarios with nuisance (van Luipen and Middelkoop, 2013). The total exposure decreased with 2.3 (Van Luipen and Schotsman, 2013).
The second researched examined the probability that a red signal will be passed, this research is described in chapter 6. The outcome of the TT-STS research is a decreased risk with a factor 2.6 (Van Luipen and Schotsman, 2013).

In Figure 34 is showed how the results can be generalized. The numbers in the bottom give an exact outcome of the study, and therefore are not easy generalizable. The outcome is less precise if an interval is given to the outcomes, but it is better generalizable. The highest level of abstraction of the numbers is given in the top. Here is stated that the risk does not go up, this gives the highest level of generalizability to the outcomes (Van Luipen and Schotsman, 2013).

**Figure 34:** Relation between generalizability and precision of outcomes (Van Luipen and Schotsman, 2013).

Since the DSSU research only tested four different routes, is this method used in the DSSU research to accept the risk for the entire emplacement in Utrecht (Van Luipen and Schotsman, 2013). This could also be used for generalization of the results to other emplacements in the Netherlands, but only if the circumstances are similar. New research is needed if different elements are present at the other locations. This can be determined via the transferability method. Therefore, this method can be used to apply transferability quicker; transferability is still possible if elements are not exactly the same and the precision requirements of the outcome is not very strict.

### 7.6 Usability of the methods

One of the previously described methods for generalizing the results should be chosen. G-theory is a very mathematical approach. This method gives the most detailed answer on how different facets have an influence on the outcome of the study. Random samples are needed to calculate the generalizability of the facets. To accomplish this several scenarios are needed, or multiple separate areas in which different samples can be taken. Since the SA is measured in the full scope simulator that cannot be easily adjusted, not a lot of different samples can be created. Therefore this method might not be very usable in practice. The SA research of the DSSU project was too limited to use this method.
The analytical generalization of the result is possible with the DSSU project. Ideally, a theory should have been developed before the SA research was executed. In order to use analytical generalization, more research should be conducted, because the results should be checked in other research.

Case-to-case transfer is the only method in which the results of DSSU can directly be applied at other locations that have the same settings. Transferability can be applied easier if the precision requirement of the outcome is not very strict. Extra research is needed for the use of the results in non-comparable locations, in which the missing settings should be investigated.

### 7.7 Improving generalization

In order to enhance the ability of generalization of a research study a couple of rules should be taken into account. First, researchers should give the boundaries of the validity of the knowledge claims (Firestone, 1993). Lincoln et al. (1985) state that researchers should provide sufficient information about the context in which a research is carried out, other researchers can judge if the results of this research could be transferred to their research. The similarity between two settings should be defined as the degree of congruence between them (1985). If research settings are typified it could make the transferability much better (Lincoln et al., 1985).

Polit et al. (2010) gives a couple of possible strategies to enhance generalizability:

1. Replication of sampling
2. Replication of studies
3. Thinking conceptually and reflexively
4. Know thy data
5. Thick description
6. Mixed method research

Replicating samples and studies make a study more robust. New studies can be designed by changing the conditions, context, and populations of the study. When the conclusions of the study results still stand, the robustness of the findings are increased (Polit and Beck, 2010), because the likeliness of wrong conclusions is reduced.

Better insights to the generalizability of a study can be obtained by thinking conceptually and reflexively about it. During a study more is known about the setting, participants, and data. Relevant information for generalization might be overlooked during the design of the study. Developing a strong conceptual and theoretical basis can enhance analytical generalization (Polit and Beck, 2010). The theory can then be checked in the study.

The researcher should thoroughly understand their data. Only when the data is fully understand, correct generalizations can be made. Misinterpretations of the data can lead to wrong conclusions.

The researcher will know the research conditions best; the conditions of the receiving party will be unknown. By describing the conditions very thoroughly (thick description), other researchers can decide if the results also apply to them. The transferability can also be made easier by applying the
proximal similarity model (Campbell, 1986). In this model the theoretical perspective to which context the findings might be transferable, is described. A gradient of similarity can then be given to the people, settings and context. Readers can then make theoretical informed decisions about the applicability of the results to their own setting (Polit and Beck, 2010). The same principle is also described by Lincoln and Guba (1985) as the level of “fittingness”.

Qualitative and quantitative data can be combined in a mixed method research. The large sample of the quantitative research can give confidence to the results of the detailed knowledge of the qualitative research (Polit and Beck, 2010).

7.8 Elements for generalization

To enhance generalization of the DSSU project, some factors that influence the outcome and therefore the generalizability can be identified. The elements can be divided in two types: static and dynamic elements. Static elements are elements that do not change frequently such as layout of the tracks, allowed static section speed, and stations. Dynamic elements change much more often such as the timetable, type of train, and type of train succession. Dynamic elements can function as prerequisites of the generalization of the static elements. For example, a study can determine that the results can be generalized over all two-track infrastructures with a visibility of 800 meters given the type of train and the train in front of it. The elements are also divided in train and track conditions, and external conditions. The train and track conditions are conditions that can be (partly) influenced by the ProRail. External conditions cannot be influenced, but might sometimes be very important for the application conditions of new designs and should therefore also be tested.

For every new design should be investigated if these elements, or combination of elements, will influence the outcome of the study. It can be used for changes at the infrastructure that are not covered in the technical system design (in Dutch: seintechnisch ontwerp) or changes in the cabin, such as new or extra devices in the train. For example, the monotony of the route will probably not influence the safety of a new sort of shunting movement. If in a study not all the essential elements were included, elements can be identified which were in the study. Then case-to-case transfer is possible for all situations where the same elements occur. In Table 6 all important sections elements for generalization are given. This list was made by brainstorming with help of three ProRail employees. First, the manager of the DSSU research project, made the initial set-up of the list. I updated this list based on experiences of the DSSU project and brainstorming about it. When this list was made an interlocking specialist, partly responsible for making OVS (in Dutch: ontwerpioorschriften, in English: designs requirements), spend an hour to check the list for relevant and irrelevant elements. After that, he updated the list with missing elements. Finally, a railverkeerskundige, expert of the rail system, brainstormed for half an hour about the elements without looking at the list. Then he checked the list for relevant and irrelevant elements and updated it based on the items he saw. This list should be updated when new insights are gathered.

The static elements are divided in four categories: signs, tracks, safety systems, and monotony of route. For the signs, both static plates as signals that give light are included. A lot of variety in placement and the displayed signal can be made. A lot of the track characteristics are already...
included in the signs since these signs are placed on specific track points; only the amount of parallel tracks is not included.

For some studies the safety systems are also important, the systems have characteristics that might give cognitive problems on the tracks. For example, if a train is equipped with ERTMS the train driver does not have to look at the green and yellow lights anymore. A train driver might be confused if the lights at the tracks give different information than the ERTMS system.

Table 6: Static and dynamic elements for generalization

<table>
<thead>
<tr>
<th>Static/track conditions</th>
<th>Static</th>
<th>Dynamic</th>
<th>Type of train succession:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signs:</td>
<td></td>
<td></td>
<td>- At station:</td>
</tr>
<tr>
<td>Visibility in amount</td>
<td></td>
<td></td>
<td>Departure: D-A,</td>
</tr>
<tr>
<td>Static section speed</td>
<td></td>
<td></td>
<td>Non-stop: N-A, N-D, N-N</td>
</tr>
<tr>
<td>Amount of signs along the track</td>
<td></td>
<td></td>
<td>- Train type Train 1 – Train 2</td>
</tr>
<tr>
<td>Evenness in section sizes (in distance)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fixation (poles, portals, dwarf signals)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tracks:</td>
<td></td>
<td></td>
<td>- Train:</td>
</tr>
<tr>
<td>Number of parallel tracks</td>
<td></td>
<td></td>
<td>- Type (freight / ICE / Intercity / sprinter)</td>
</tr>
<tr>
<td>Safety systems:</td>
<td></td>
<td></td>
<td>- Type (NS / Arriva / Veolia etc.)</td>
</tr>
<tr>
<td>- ATB-EG / NG / VV</td>
<td></td>
<td></td>
<td>- Type (ICM / SLT / etc.)</td>
</tr>
<tr>
<td>- ERTMS Level 1 / 2 / 3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Disturbed/unusual situations: |
| - Delay                      |
| - Different train movements (backwards, train on other track) |
| - Shunting movement          |
| - Unexpected situations (tree on electricity tracks, cow on rail, disrupted crossings etc.) |
| - Maintenance                |

Activity on the track:
- Crowded / quiet

<table>
<thead>
<tr>
<th>External conditions</th>
<th>Monotony of route:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>- Amount of stations</td>
</tr>
<tr>
<td></td>
<td>- Bridges</td>
</tr>
<tr>
<td></td>
<td>- Tunnels</td>
</tr>
<tr>
<td></td>
<td>- Switches</td>
</tr>
<tr>
<td></td>
<td>- City / countryside</td>
</tr>
</tbody>
</table>

| Environment conditions: |
| - Night / day          |
| - Weather conditions   |

Train driver:
- Experience level
- Habitation
- Age
- Mood (stressed, tired etc.)
- Time worked in that shift
- Distraction factors (extra devices in train, telephone etc.)
Finally the monotony of the route is another element. These elements can often not be influenced by ProRail. The monotony can have an influence on the driver’s performance; different elements in the route make the route less monotone.

The dynamic elements can be divided in: type of train succession, type of train, disturbed situations, activity on the track, environment conditions, and train driver characteristics. Two trains can have nine different train successions at stations. These successions have an influence on the signs displayed and the TT-STS of trains.

The type of train can also have an influence on the outcome of the study, often a difference in freight, intercity or sprinter will suffice. These trains can get different signs displayed; an intercity behind a sprinter (slower train) might get more yellow signs than when it is following another intercity.

Some other situations will need more detailed train information. The disturbed situations have different impacts. A delay will give more yellow signs. Sometimes other routes should be chosen or the direction of drive should be changed if there is an obstruction on the track, this can lead to dangerous situations if this is not tested or thought about.

The activity on the track has a direct influence on the displayed signs. These elements can all be influenced in some extend by ProRail.

Environment conditions cannot be influenced, but might have an influence on the outcome of the study. For example, when the weather is foggy, the signs are only visible at the moment of passing the signal, numbers on the signs might be missed by train drivers. Finally, the train driver’s style of driving has a lot of influence on the outcome of the study. Although some requirements can be set to train drivers, a lot of these factors cannot be influenced by ProRail. Habituation of the train driver is the most difficult to measure. For example, if a sign normally always shows yellow + a blinking number and in one situation it just shows yellow, the train driver might miss this, which can have disastrous results.

### 7.9 Generalizability of the results

The SA research was based on one specific route for DSSU, therefore the outcome only holds for the situation in Utrecht. The scenario designs were optimised for normal conditions during a normal sunny day. Unfortunately, generalization of the results was not taken into account during the study design; management gave priority to the specific results for DSSU. All elements that influence generalizability are identified in chapter 7.8, in this chapter the list of missing elements can be made to generalize the results. Therefore this chapter contributes to an experimental set-up for DDDR in the Netherlands.
In Table 7 are the elements again shown and is indicated which elements were included and varied in the DSSU studied route; elements included are indicated with a ✔, non-included elements are indicated with a ✖.

For the static train or track elements, all the sign and track elements were varied in the route. The study was executed with the ATB system; ERTMS was not included since this is not available on this track.

For the dynamic train or track elements, not all the types of train successions are included. The SA measurement only included the route from Culemborg to Utrecht; only the second train stopped at Utrecht. The train driver drove an intercity train that followed a freight train. Half of these intercity trips were made in a SLT train, the other half in a VIRM. There were two different scenarios: with and without nuisance. This nuisance gave a delay and can also be seen as a crowded versus quiet track.

The only element for static external conditions is the monotony of the route. The driven route was not monotonous; three stations, two bridges, lots of switches were passed on the route, and both cities and countryside were included.

The dynamic external conditions exist of environment conditions, which were not included, and train driver characteristics, which were partly included. The only elements that were included were age and experience level.

This SA research was part of a bigger research in which more elements were included. The same route with the same characteristics of the tracks was used for the eye-tracker and hart rate measurements. More elements were included in the TT-STS. More types of train succession at station were included: Arrival - Arrival, Arrival - Departure, Departure - Departure - Non-stop, and Non-stop - Non-stop. Also more different trains between the different routes were used: Sprinter-intercity, intercity-intercity, intercity-ICE.
After the DSSU research was finished, all the internal stakeholders from ProRail and NS were invited for a presentation in which all the researchers could answer questions about the research. Some elements were identified that needed further research before the results can be generalized. ‘Treinbeveiligers’ (Employees from the train safety department) identified some elements that they wanted to be investigated further: a departure-departure train succession at the station with a sprinter – intercity, and freight- and student drivers should be included in the study.

<table>
<thead>
<tr>
<th>Table 7: Presence of elements for generalization</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Train / track conditions</strong></td>
</tr>
<tr>
<td>Static</td>
</tr>
<tr>
<td>Signs:</td>
</tr>
<tr>
<td>✓ Visibility in distance</td>
</tr>
<tr>
<td>✓ Visibility in amount</td>
</tr>
<tr>
<td>✓ Static section speed</td>
</tr>
<tr>
<td>✓ Amount of signs along the track per distance</td>
</tr>
<tr>
<td>✓ Evenness in section sizes (in distance)</td>
</tr>
<tr>
<td>✓ Fixation (poles, portals, dwarf signals)</td>
</tr>
<tr>
<td>Tracks:</td>
</tr>
<tr>
<td>✓ Number of parallel tracks</td>
</tr>
<tr>
<td>Safety systems:</td>
</tr>
<tr>
<td>✓ ATB-EG / NG / VV</td>
</tr>
<tr>
<td>✗ ERTMS Level 1 / 2 / 3</td>
</tr>
<tr>
<td>Dynamic</td>
</tr>
<tr>
<td>Type of train succession:</td>
</tr>
<tr>
<td>- At station:</td>
</tr>
<tr>
<td>Arrival</td>
</tr>
<tr>
<td>Arrival</td>
</tr>
<tr>
<td>Departure</td>
</tr>
<tr>
<td>Non-stop</td>
</tr>
<tr>
<td>✓ Train type Train 1 – Train 2</td>
</tr>
<tr>
<td>Train:</td>
</tr>
<tr>
<td>✗ Type (freight / ICE / Intercity / sprinter)</td>
</tr>
<tr>
<td>✗ Type (NS / Arriva / Veolia etc.)</td>
</tr>
<tr>
<td>✓ Type (ICM / SLT / etc.)</td>
</tr>
<tr>
<td>Disturbed/ unusual situations:</td>
</tr>
<tr>
<td>✓ Delay</td>
</tr>
<tr>
<td>✗ Different train movements (backwards, train on other track)</td>
</tr>
<tr>
<td>✗ Shunting movement</td>
</tr>
<tr>
<td>✗ Unexpected situations (tree on electricity tracks, cow on rail, disrupted crossings etc.)</td>
</tr>
<tr>
<td>✗ Maintenance</td>
</tr>
<tr>
<td>Activity on the track:</td>
</tr>
<tr>
<td>✓ Crowded / quiet</td>
</tr>
<tr>
<td>External conditions</td>
</tr>
<tr>
<td>Monotony of route:</td>
</tr>
<tr>
<td>✓ Amount of stations</td>
</tr>
<tr>
<td>✓ Bridges</td>
</tr>
<tr>
<td>✗ Tunnels</td>
</tr>
<tr>
<td>✓ Switches</td>
</tr>
<tr>
<td>✓ City / countryside</td>
</tr>
<tr>
<td>Environment conditions:</td>
</tr>
<tr>
<td>✗ Night / day</td>
</tr>
<tr>
<td>✗ Weather conditions</td>
</tr>
<tr>
<td>Train driver:</td>
</tr>
<tr>
<td>✓ Experience level</td>
</tr>
<tr>
<td>✗ Habituation</td>
</tr>
<tr>
<td>✓ Age</td>
</tr>
<tr>
<td>✗ Mood (stressed, tired etc.)</td>
</tr>
<tr>
<td>✗ Time worked in that shift</td>
</tr>
<tr>
<td>✗ Distraction factors (extra devices in train, telephone etc.)</td>
</tr>
</tbody>
</table>
The disturbed/ unusual situations, day and night situations and weather conditions were not included in the study. The study design is optimised for normal conditions. Train drivers are expected to act safely by lowering speed when dangerous situations occur.

Other factors are very difficult to include in the study. A monotone route can be created, but train drivers will be extra alert in a simulator because they know they are tested. Habituation is also very difficult to include. Habituation can lead to dangerous situations: a train driver can easily miss a yellow sign when this sign is in normal condition showing yellow with a blinking number, but the meaning of the signs differ (be prepared to stop at next sign versus the speed limit of the blinking number cannot be reached at the next sign. Braking should not be interrupted if next sign allows a lower speed than your speed). This situation is very hard to include in the simulator, since these effects only occur when a train driver is very well known with the situation and should therefore have driven this situation very often.

Overall, only the train succession departure-departure and freight- and student drivers should be added to the study, to generalize the results. These are the only two missing factors that were identified by experts as important for the study.

### 7.10 Reflection

Ideally, the missing elements that were identified in chapter 7.7 would have been implemented in the initial measurements as described in chapter 6. An important condition for doing this is that the results of the DSSU study should not have a reduced reliability. Project management stated that because of the importance of DSSU the research should be aimed specifically at DSSU. Adding extra elements to the scenarios, which do not apply to DSSU but only to DDR, makes the scenarios less focused. It is very hard to draw conclusions on DSSU if negative study results were found in scenarios that are not applicable to the DSSU project. The simplest way to do this is by adding extra scenarios that are tested after the DSSU scenarios, but the practice showed that this is not so easy. Making scenarios in the full scope simulator was very hard. The four tested scenarios were only available 0,5 hour before the tests began.

Not a lot of elements were varied in the SA measurement, since this was not possible in the full scope simulator. More variation could have been implemented in the TT-STS study, since the SA measurements are used as a validation of the TT-STS study. For the determination of the SA a worst-case scenario can be developed, as illustrated in Figure 36. A worst-case scenario, in normal conditions, for the SA of train drivers can be selected from the identified elements in chapter 7.8. Experts should determine this worst-case scenario. This scenario can be used if it applies to DSSU; the result is an outcome that is both applicable to the project as to a general system-level change. If it does not apply to DSSU, management has to make the decision whether the worst-case scenario will be tested for SA or that the SA measurement will not be applicable to generalize the results. The advantage of testing a generalized scenario is that no extra research is needed after the project. The disadvantage can be a negative study outcome: no conclusions can be drawn on the project when the scenario is not applicable to the project and the outcome of a generalized result is negative.
7.11 Consideration of strategic management

Strategic management has decided that DDDR will not be further investigated at this moment; DDDR can be applied to other situations as well. They give three reasons:

1. A monitor program will be started to monitor DSSU + DDDR (Van Luipen and Schotsman, 2013).

When this project is executed a monitor program will be started, for monitoring any safety issues.

- The amount of red signal approaches will be monitored.
- Habituation of train drivers will be monitored by measuring the accelerating and braking behaviour.
- Fatigue of train drivers will be monitored because of repetitive DDDR.
- Transition mistakes of train drivers will be monitored because of driving on routes with and without DDDR (Van Luipen and Schotsman, 2013).

By monitoring these data any dangerous non-researched behaviour can be identified. An intervention can be executed when dangerous situations seem to occur before an actual SPAD occurs.

2. None of the pre-determined hazards were found to be present in the DSSU situation (van Luipen and Middelkoop, 2013).
A HAZID (hazard identification) study was executed before this research was done. In this HAZID study some hazards with their causes were identified specifically for the DSSU project. From this study some application conditions of DDRD were set-up and implemented before the measurements were executed. Thereafter, none of the identified causes occurred dominantly during the research. For this reason no restrictions are set for the amount of blocks on which a braking occurs and the number of DDRD in sequence (van Luipen and Middelkoop, 2013).

3. The results of the study did not show any significant differences in SA between the current and the DSSU situation (Van Luipen and Schotsman, 2013).

The outcome of the study showed no significant differences between the different situations. Strategic management expects that it is very likely that no significant differences will be found in the extra non-researched situations.

### 7.12 Comments

Of course there are some pitfalls and difficulties for generalization. Firstly, in research the settings of the research facets are dichotomous. In real life there is a continuum between settings (Rodon and Sesé, 2008). Here for the real situation can never be exactly modelled, choosing the levels of the settings can be very hard. Cronbach stated: “When we give proper weight to local conditions, any generalization is a working hypothesis, not a conclusion” (Cronbach, 1972, p. 125). Therefore, these generalization conclusions have to be tested in new studies, since generalizations do not apply to particulars (Lincoln et al., 1985).

### 7.13 Conclusion

This chapter discusses the possibilities for generalization of the SA study results. Three possibilities were discussed: generalizability theory, analytical generalization, and transferability or case-to-case transfer.

Generalizability theory is used to determine the dependability of an observation, or how accurately a specific, observed sample of behaviour reflects actual behaviour under the range of all possible facets of the measurement. A decision study is used for determining the set of conditions to which a decision maker wants to generalize.

In analytical generalization the results are generalized to a broader theory. Analytic generalization is facilitated by specifying the conditions under which a study is done and their relevance to multiple theories. That knowledge is used to create controls, quasi-experimental designs, or replications that strengthen generalization.

Transferability or case-to-case transfer is used to apply the results of one setting to another. It can be enhanced by thick description that allows assessment of the applicability of study conclusions to
one’s own situation. Transferability can be applied easier if the precision requirement of the outcome is not very strict.

Ideally, generalizability is taken into account before a study is executed. The study design can then be adjusted to generalize to all conditions that are applicable. Generalizing to a broader theory is then preferred via analytical generalization or generalizability theory. Case-to-case transfer can be used when the study design is not on beforehand designed with generalizability taken into account.

Important elements for the generalizability of the study are identified in static and dynamic elements. The dynamic elements are the prerequisites for the applicability of the generalization of the static elements. These elements can be used to design a study on beforehand or identify elements in case-to-case transfer. Not all elements that could have an influence on the outcome of the study were included, but strategic management has decided that no further research is needed at this moment. DDDR can be applied in other places as well. A monitor program will be started to monitor safety issues in order to intervene when dangerous situations occurs.

In future research that has to be generalized, experts can identify the worst-case scenario, in normal operating conditions, for SA. If this scenario is part of the project it can be used to measure the SA. If the scenario is not part of the project, management has to decide whether they want to measure only the scenario that is applicable to the project or that the measurements cannot be generalized. The advantage of testing a generalized scenario is that no extra research is needed after the project. The disadvantage can be a negative study outcome: no conclusions can be drawn on the project when the scenario is not applicable to the project and the outcome of a generalized result is negative.
Decision-making based on measurements of behavioural safety
8 Conclusion

In this chapter the conclusions and recommendations will be given. The conclusions will be given based on the research questions as stated in the introduction.

8.1 Conclusion

The goal of this thesis was to research how decisions can be made upon measurements of behavioural safety. ProRail wants to be able to measure changed behaviour of train drivers to test the safety of new designs on the rail or in the train cabin. Decisions should be made based on these measurements by the project management, who are primarily interested in the results for their project, and strategic management, who are interested in if the results also apply to other settings.

Situation Awareness (SA) was used to measure the changed behaviour, because previous research suggested that Dynamic Double non-interrupted braking (DDDR) might have negative consequences for the expectations of the train driver. SA is used in this thesis to measure the behaviour of train drivers, because DDR was the inducement for this thesis. The main research questions which is answered is therefore:

How to measure SA of train drivers in order to compare the safety of different infrastructure designs, taking into account that both project and strategic management should be able to make decisions based on these measurements?

To be answer the research question, first the sub research questions should be answered. A thorough understanding of SA is needed to understand how safety can be compared based on these measurements. Therefore the first sub research questions is:

1. What is SA and how is it related to safety?

The definition of SA that is used in this thesis is:

‘the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future’

(Endsley, 1988a, p. 97)

A person forms his SA based on the elements in the environment and their status. Thereafter a decision can be made based on this SA. If the situation changes by this decision, the SA is updated and the cycle begins again. A lot of factors have an influence on the SA. Long-term memory helps to recognize situations; this is memorized in the mental model of a person. The goals and objectives of a person help to focus on the important elements in the environment. The expectations and automaticity can both help and hinder for gaining SA.

Although there is no consensus about the exact relations between SA and safety, all theories agree that SA is a requirement for making informed, conscious decisions. Of course more factors influence safety such as skills and properly working equipment. SA is therefore one of the many elements that influence safety.
In order to be able to make decisions on the SA measurements, the requirements of the acceptability of the results are researched in sub question 2 and 3.

2. What are the requirements of decision-makers to accept the results of an SA study for the project?

Three factors are identified that have an influence on the acceptability of the results. First the safety culture ladder; negative results will be easier accepted in a company that is high on the safety culture ladder. The safety culture ladder is divided in a pathological, reactive, calculative, proactive, and generative culture, in which the pathological culture is the lowest safety culture and the generative culture is the highest. Examples of a reactive and calculative culture were found. A company with a reactive culture investigates safety issues after an accident occurred; safety is approached via rules and regulations. Calculative companies only take safety issues in account if improvements are proven to work. ProRail is not neglecting safety issues; it is likely that clearly negative study outcomes will be accepted. There will be a danger in outcomes that are not clearly very bad; economic reasons might get dominant in the decisions they make. The DSSU-project (DoorStroomStation Utrecht) is the first step for ProRail into a more proactive safety culture; in this project the safety is first tested before it was implemented.

Second, strategic decision-making can have a big impact on the acceptability of the results. The decision-makers are the people who have to accept the results of the study. Three steps of decision-making are identified: goal formulation, alternative generation, and evaluation and selection. In all these steps are possible threats for the acceptability of the results. Other interest might be a deciding factor if safety is not incorporated in all these steps. Safety will be incorporated in all steps if a company has a high safety culture. In the DSSU project the biggest danger was that no alternatives were left when the safety research was executed, stopping the project was therefore very hard. It will be easier to stop a project if a stage-gate model is used in which every stage of the project would have been checked on safety. In order to enhance the acceptability of the results, the decision-makers should be kept closely informed about the measurements, and letting them decide on all the design choices that had an impact on the uncertainty and robustness of the study. It will be difficult for decision-makers to reject the results if they personally agreed to all the choices.

The third, more formal, requirement of the acceptability of the results is the use of the Common Safety Method (CSM), which is used by ProRail. For the measurements there are three requirements: the safety should be at least as good as reference systems, the methods used should reflect the system correctly under assessment and its parameters, and the results must be sufficiently accurate to serve as robust decision support.

3. What are the requirements for an SA study design for train drivers?

For the measurements a comparison should be made between the new design and a reference system. The best method for comparing designs is a controlled before and after study. In that study the situation before an intervention is compared with the situation after an intervention. Multiple before and after studies, a big magnitude of change, and a logically linked intervention to the change in outcome, will make the study more robust. The best control on the before and after studies is
obtained by using a simulator; the researcher can then exactly determine which elements should be changed in the study.

People are being measured in SA studies, when people are aware of the measurements they might show different behaviour than they normally do. Since measuring SA is impossible without the help of the participants, it is key that they do not know when the SA is tested. The design should be made in a way that at random moments the SA is measured, in order that the participant cannot prepare for the measurements.

The end points of the study are the points of measurements. It should be valid and reliable; it should measure what is stated that it will measure. If the measurements are repeated it should give the same values. If end points are not available, surrogate endpoints can be chosen. The surrogate endpoint should be able to predict the outcome of the study.

A mixed method research makes the outcome more robust. In a mixed method design quantitative research is combined with qualitative research. If different methods all have the same outcome, the conclusion is more robust. If the outcomes are not the same, the reason can be detected by combining the outcomes and a better understanding of the outcome is gained.

All the requirements for decision-makers to accept the outcomes of an SA study are given in sub-question 2 and 3, based on these requirements a measurement method or set of methods should be chosen.

4. Which SA measurement method or set of methods is best to choose based on the requirements?

The difficulty of measuring SA is that not all SA measurement methods are validated as an SA measurement method. The most important requirements for SA measurement methods are the validity, reliability, sensitivity and the lack of intrusiveness; the methods should measure what is stated that it will measure, when a measurement is repeated it should give the same values, it should be able to measure differences between designs, and it should not direct the SA of a person to certain elements. The measurements will be executed in a high fidelity simulator that can be paused during a simulation. As a basis the direct objective Situation Awareness Global Assessment Technique (SAGAT) method can be used. It is the best-known validated, reliable, and sensitive method for objective measuring SA.

In order to make the outcomes of SAGAT more reliable, it can be combined with other methods since a mixed method is preferred. A combination of SAGAT with a post-trial subjective method, an observer rating method, performance measures, and process indices are possible. The post trial subjective method Mission Awareness Rating Scale (MARS) is a quick and easy method for measuring SA, which does not require a lot of training and is not intrusive. Therefore, it can easily be combined with SAGAT. The Situation Awareness Behavioural Rating Scale (SABARS) is the best-known objective measurement method, and is suitable for measuring SA of train drivers. TT-STS (time to signal passed at danger is irrevocable) is a newly developed performance measure for train drivers, which can be combined with SAGAT without influencing the SAGAT outcome. The eye tracker can be used for measuring process indices, which can also be combined. All the methods can be combined with each other. SAGAT should be combined with at least one other method. MARS is the easiest method to
combine with SAGAT, and the measurements do not influence each other. Therefore, it is recommended to combine MARS with SAGAT.

The requirements and measurement methods are given in sub question 2, 3 and 4. The next step is to research if the SA measurements can be used to make a safety decision on in practice.

5. Are SA measurements useable to make a safety decision on?

SAGAT and MARS were used to measure the SA of the train drivers in the case study. No differences in SA were found between the old situation and the DSSU situation, but differences were found between different nuisance scenarios; this implies that the measurements were sensitive enough to identify differences. Only SAGAT was sensitive enough to identify these significant differences in SA, but MARS showed the same trends in SA based on optical comparison of MARS scores, which made the SAGAT outcome more robust. MARS might be able to show significant differences if more participants are available, but more research is needed to conclude this.

SA was usable as a supportive method for making safety decisions on. SA is used to give an insight into the brain of the train driver: it indicates how aware the train driver is of the situation and therefore was able to make informed, conscious decisions. SA cannot be used as a single method for determining the safety since SA is only a requirement, and not a guarantee, for making informed, conscious decisions. SA as a single indicator for safety can therefore only be used when the SA results are bad. SA was combined with TT-STS and workload measurements to indicate the safety. TT-STS indicates the probability of a Signal Passed at Danger (SPAD), the workload measurements gave a further insight into the changed workload of the train drivers. Since both these measurements also indicated that there is no difference between the old situation and the DSSU situation, a robust answer is given to the safety case of DSSU: DSSU is not less safe than the current situation.

The DSSU project only answered how project management can use behavioural safety measurements to make safety decisions on. The results have to be generalized in order for strategic management to make a safety decision on applying DDDR elsewhere in the Netherlands. Therefore the last sub research question is:

6. How can the results of the SA study be generalized?

Three theories describe how the results can be generalized. Generalization theory can be used to determine which facets have an influence on the outcome of the study. This is a very mathematical approach and the most detailed method. Since multiple random samples should be taken from the scenarios with different facets, multiple scenarios or areas within scenarios should be made to use this method. This was not very feasible in the simulator that was used in the experiments.

Analytical generalization can be done by developing a theory, and thereafter check the theory with measurements. More research is needed to generalize the results of the DSSU project, since the results should be checked in other researches.

Case-to-case transfer is the only generalizability theory that does not need additional research before DSSU can be applied at other locations. Elements should be identified which were included in the study. The results of DSSU can be applied at another location if the same elements are present. Case-
to-case can be applied easier if the precision requirement of the outcome is not very strict. Extra research is needed if elements are present that were not included in the DSSU study.

Important elements that influence the outcome of an SA study are identified in a list. From this list the elements can be identified that influence the SA, which can then be used to make scenarios. The results can be generalized if all these scenarios can be tested. Since SA is used to give robustness to a TT-STS study, not all scenarios have to be tested. If not all scenarios can be tested a worst-case scenario, in normal operating conditions, can be made. This scenario can be used if this scenario is part of the researched project, if this is not the case, management has to decide whether they want the results to be generalizable or if they want to have a specific scenario for the project.

All the answers given above contribute to the answering of the main research question:

How to measure SA of train drivers in order to compare the safety of different infrastructure designs, taking into account that both project and strategic management should be able to make decisions based on these measurements?

SA measurements can be used as an input for decisions on the safety of infrastructure designs when it is combined with TT-STS measurements. SA measurements as a single indicator will not suffice to make safety decisions on because the SA of train drivers is a requirement, but not a guarantee for train drivers to make informed, conscious decisions. Therefore, management can only use SA if the SA of the train drivers is incorrect. SAGAT method is useful for measuring SA of train drivers; it is sensitive enough for measuring differences between designs and gives objective and direct SA data. The outcome of the study was made more robust by combining the SAGAT method with the MARS method.

Measurements have to be applicable for the project, for the project management; and measurements should be applicable on a system-level, for strategic management; project management wants specific answers, strategic management wants generic answers. Since SA measurements should always be combined with TT-STS measurements, the SA can only be measured in the worst-case scenarios. The TT-STS measurements are used to give statistical power to the study; the SA is used to get a better insight into this data. Important elements of the rail infrastructure are identified, which can help for making worst-case scenarios for SA, in normal conditions. This scenario can be used if this scenario is applicable for the project. When it is not applicable to the project, management has to decide whether they want the results to be generalized, or select another scenario that is only applicable to that project. The advantage of testing a generalized scenario is that no extra research is needed after the project. The disadvantage can be a negative study outcome: no conclusions can be drawn on the project when the scenario is not applicable to the project and the outcome of a generalized result is negative.
8.2 Recommendations

Some recommendations can be made based on findings in this report.

Generalizability should be taken into account before the project starts. Although some projects want specific SA results for their project, often a scenario can be chosen that is both applicable to the project as is needed to generalize the results. If this is done properly, no extra research is needed after the main research.

The MARS method should be further researched as an SA measuring method for train drivers. The study of DSSU had too little participants to see any statistical differences between designs. A study that is performed with more participants can be used for researching the usability of the MARS method. MARS was also used as an observer rating method, but the sensitivity of these measurements was worse than when the train drivers used MARS. A better observer rating method should be used in new SA research; SABARS could be well suited for measuring SA for train drivers.

MATRICS was not usable for determining SA in this project, since no differences in SA were found, but small differences were found in the full scope simulator. Big differences are much more interesting to know, since then there is a clear relation between the SA and the design. When a study measures big differences in the full scope simulator these measurements should also be measured in MATRICS. A full scope simulator is no longer needed if these big differences are also found in MATRICS. This has two main benefits: MATRICS is much cheaper and easier to set up.

As an alternative yet another simulator could be used, which has more detail than MATRICS but not a complex motion driven cockpit. This simulator should be able to accurately simulate the behaviour of a train and the rail system. A detailed logging system should also be implemented, and the scenarios should be easily made. As an extra it would be an advantage if a lot of train and track failures could occur in the scenarios, these events can be tested for SA measurements.

In this thesis a list with important elements for generalization is identified. This list should be constantly updated when new insights are gathered from either accidents or new brainstorm sessions.

A stage-gate model can be implemented in the project structure at ProRail. More moments of reflection on the safety of the project are available when this model is implemented. Decision-loops are shorter which helps to make solve the escalating commitment problem.
9 Reflection

In this chapter I will reflect on my thesis. Three questions will be answered:

- How did my thesis contribute to current research?
- How useful is measuring SA of train drivers?
- What did I learned from my thesis?

9.1 How did my research contribute to current research?

This project was the first project in which different infrastructure designs for trains are tested with SA measurements. My research showed that measuring SA is possible via the SAGAT method and differences between designs can be measured. Furthermore, elements are identified which can be used to check whether they might have an impact on the SA of a train driver in a new design. Most of these elements can be used worldwide for developing SA studies of train drivers.

In this thesis a comparison was made between the SAGAT and MARS outcomes, which was recommended by Matthews et al. (2002). Although not sufficient data was available to make a statistical comparison between the methods, the study identified the same trends in outcome.

SA as a single method for determining safety of a system is not enough, since people can make mistakes while having a correct SA. Combining SA with a qualitative performance research gives a detailed insight into safety issues of a design; qualitative research gives power to the outcome, the SA gives insight in why the outcome is like it is.

9.2 How useful is measuring SA of train driver?

Situation awareness is often used in military and aviation. The situations they are working in, is often very complex. A jetfighter pilot for example has to pay attention to his altitude, speed, locations of allies, location of enemies etc. A train driver should primarily take into account his speed and the signals. His main movement is giving traction and brake. Of course a train driver also has to look at other things, such as dangerous situations on the tracks, but these are relatively easy to identify. In the DSSU study not a lot of significant differences were found. If future research also identifies no significant differences in SA, it might be recommendable not to measure SA anymore.

9.3 What did I learned from my thesis?

Apart from all the theories which are discussed in this project, I also learned a lot of the process of making the thesis. In the beginning of my thesis I rushed into measuring SA. The measurements were scheduled in the beginning of the thesis, and therefore I spent all my time in reading about SA,
reading about the measurement methods, and consult some experts. I ‘forgot’ about my thesis till the measurements were done. I did not know what to do next, since all the measurements were already done with a lot of requirements taken into account. I had the unnecessary feeling that I had to produce all the ideas of the thesis by myself in which the committee would correct me if I were wrong. After I started asking them more questions about content and structure new leaps were made in my thesis. I had lots of ideas but did not know how to structure it or where to set the limits of the thesis. With help of the exam committee I structured my report and researched more theories about study design and decision-making. The outcome is that I can now present this scientifically supported report.
Appendix

Appendix A: Glossary

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>English</th>
<th>Dutch</th>
</tr>
</thead>
<tbody>
<tr>
<td>CARS</td>
<td>Crew awareness rating scale</td>
<td>Crew awareness rating scale</td>
</tr>
<tr>
<td>DDRR</td>
<td>Dynamic double non-interrupted braking</td>
<td>Dynamisch dubbele doorgaande remming</td>
</tr>
<tr>
<td>ERTMS</td>
<td>European rail traffic management system</td>
<td>Europees spoorwegbesturingssysteem</td>
</tr>
<tr>
<td>GDTA</td>
<td>Goal directed task analysis</td>
<td>Taakgerichted analyse van taken</td>
</tr>
<tr>
<td>HAZID</td>
<td>Hazard Identification</td>
<td>Gevaarlijke identificatie</td>
</tr>
<tr>
<td>HRO</td>
<td>High reliability organisations</td>
<td>Hoog betrouwbare organisaties</td>
</tr>
<tr>
<td>HTA</td>
<td>Hierarchical Task Analysis</td>
<td>Hiërarchische takenanalyse</td>
</tr>
<tr>
<td>MARS</td>
<td>Mission awareness rating scale</td>
<td>Missies awareness rating scale</td>
</tr>
<tr>
<td>NSO</td>
<td>NS Dutch Railways educations</td>
<td>NS opleidingen</td>
</tr>
<tr>
<td>SABARS</td>
<td>Situation awareness behavioral rating scale</td>
<td>Situation awareness behaviordrating scale</td>
</tr>
<tr>
<td>SAGAT</td>
<td>Situation awareness global assessment technique</td>
<td>Situation awareness globaal beoordelingstechniek</td>
</tr>
<tr>
<td>SART</td>
<td>Situation awareness rating technique</td>
<td>Situatie awareness rating technique</td>
</tr>
<tr>
<td>SLT</td>
<td>Sprinter lighttrain (type of train)</td>
<td>Sprinter Lighttrain (type trein)</td>
</tr>
<tr>
<td>SME</td>
<td>Subject matter expert</td>
<td>Subject matter expert</td>
</tr>
<tr>
<td>SPAD</td>
<td>Signal passed at danger</td>
<td>Signaal overgegaan als gevaar</td>
</tr>
<tr>
<td>TT-STS</td>
<td>Time to SPAD</td>
<td>Tijd tot passage stoptonend sein</td>
</tr>
<tr>
<td>VIRM</td>
<td>Type of train</td>
<td>Type van train</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix B: List of figures

Figure 1: Structure of the project .................................................................................. 10
Figure 2: Structure of the report .................................................................................. 11
Figure 3: This chapter contributes to the project problem definition ......................... 13
Figure 4: 3 Levels of SA (Endsley et al., 2003) ......................................................... 14
Figure 5: Mental model schema (Jones and Endsley, 2000) ........................................ 15
Figure 6: Interactive sub-system (Bedny and Meister, 1999) ...................................... 18
Figure 7: Neisser’s (1976) perceptual cycle, with the invariant in the core (Smith and Hancock, 1995) ................................................................. 18
Figure 8: Model of situation awareness in dynamic decision making (Endsley, 1995b) ........................................................................................................... 20
Figure 9: This chapter contributes to the project problem definition ......................... 21
Figure 10: Check for significant change in system (European Railway Agency, 2009b) .................................................................................................................. 28
Figure 11: CSM scheme (European Railway Agency, 2009b) ......................................... 31
Figure 12: This chapter contributes to the project experimental set-up ....................... 33
Figure 13: This chapter contributes to the experimental set-up .................................... 39
Figure 14: Time to STS (Van Luijen and Middelkoop, 2013) ....................................... 46
Figure 15: This chapter explains the executed SA tests ................................................. 49
Figure 16: Infrastructure layout 2013 (Van Luijen and Middelkoop, 2013) ................... 50
Figure 17: Infrastructure layout DSSU (Van Luijen and Schotsman, 2013) ................. 50
Figure 18: Different signal successions (Zon et al., 2013) ........................................... 51
Figure 19: Interior of the SLT Full Scope Simulator ..................................................... 52
Figure 20: Comparison between old infrastructure of all train drivers combined ........ 55
Figure 21: Comparison between old infrastructure and new infrastructure per train driver ........................................................................................................... 55
Figure 22: Comparison between nuisance of all the train drivers combined .......... 56
Figure 23: Comparison between nuisances per signalling design of all the train drivers combined ................................................................. 56
Figure 24: MARS question 1 ....................................................................................... 59
Figure 25: MARS question 2 ....................................................................................... 59
Figure 26: MARS question 3 ....................................................................................... 60
Figure 27: MARS question 4 ....................................................................................... 60
Figure 28: MARS question 5 ....................................................................................... 61
Figure 29: MARS question 6 ....................................................................................... 61
Figure 30: MATRICS simulation ................................................................................ 61
Figure 31: This chapter contributes to the conclusions of the project ......................... 65
Figure 32: Probability of a SPAD ............................................................................... 65
Figure 33: This chapter contributes to the goal and problem definition of DDR .......... 67
Figure 34: Relation between generalizability and precision of outcomes (Van Luijen and Schotsman, 2013) ................................................................. 71
Figure 35: This chapter explains how DDR can be tested for the Netherlands ............. 76
Figure 36: Method for determining SA scenario ........................................................... 79
Appendix C: List of Tables

TABLE 1: OVERVIEW OF RESEARCH QUESTIONS AND THEIR CHAPTER NUMBERS
......................................................................................................................... 9
TABLE 2: SUMMARY OF THE ROLE AND INPUTS OF THE FUNCTION BLOCKS (STANTON ET AL., 2001)
................................................................................................................................. 17
TABLE 3: OVERVIEW OF SA MEASUREMENT METHODS
......................................................................................................................................................... 40
TABLE 4: COMPARISON BETWEEN AMOUNT OF TIMES QUESTION ANSWERED WRONG WITH NUISANCE AND WITHOUT NUISANCE
.......................................................................................................................................................... 57
TABLE 5: MARS OUTCOME ......................................................................................................................... 58
TABLE 6: STATIC AND DYNAMIC ELEMENTS FOR GENERALIZATION
..................................................................................................................................................................... 74
TABLE 7: PRESENCE OF ELEMENTS FOR GENERALIZATION
..................................................................................................................................................................... 77
Appendix D: SAGAT question forms DSSU

Driver: ______ Scenario: _______  Driver/Observer  Date ___/5/2013  Page 1A

Circle the right answer!

1. Is the train delayed by more than three minutes?
   Yes  No

2. What was your previous signal?
   Green, green blinking + number, green blinking,
   Yellow blinking + number, yellow number, yellow, yellow blinking, red

3. Is the dead man’s switch functioning properly?
   Yes  No

4. In which type of train do you drive?
   DM’90, ICM, E-loc 1700, VIRM, SGMm, Mat ’64, mDDM,
   DDAR, SLT, DDZ, ICMm

5. In comparison with the allowed speed, your speed is:
   Too high  Good

6. Are you braking?
   Yes  No

7. Is the air conditioning functioning properly?
   Yes  No
8. How would you rate your alertness in the last five minutes on a scale?

1 2 3 4 5 6 7

extremely sleepy extremely alert

9. What do you expect that the coming signals will allow you?:

I can speed up I can keep this speed I will have to break

10. What color did the rabbit have?

black white brown blue green pink there were no rabbits

11. Are the cabin lights functioning?

Yes No

12. What is the signal behind the signal you are approaching?

Green, green blinking + number, green blinking,

Yellow blinking + number, yellow number, yellow, yellow blinking, red

13. How many bridges did you pass?

0 1 2 3 4 5

14. Are the wipers functioning?

Yes No

15. Is the cabin chair broken?

Yes No
1. Are the signals working properly?
Yes
No

2. What is your current speed? (Put an X at your speed.)

0 40 60 80 100 130 140

3. Are you driving your usual route?
Yes
No

4. Are you giving traction?
Yes
No

5. Are you fulfilling your breaking criterion?
Yes
No

6. Is de loud-speaker installation functioning?
Yes
No

7. What signal are you approaching?
Green, green blinking +number, green blinking,
Yellow blinking + number, yellow number, yellow, yellow blinking, red

8. Is all the train equipment functioning properly?
Yes
No
9. What is the speed of the train driving in front of you?:
Is driving slower than my train
Is driving the same speed
Is driving faster than my train
There is no train driving in front of me

10. Were there any problems with the GSM-R?
Yes   No

11. What type of train is stationed at OZ?
DM'90,   ICM,   E-loc 1700,   VIRM,   SGMm,   Mat '64,   mDDM,
DDAR,   SLT,   DDZ,   ICMm

12. Did you pass any defect level crossings?
Yes   No

13. Is the typhoon functioning?
Yes   No

14. Did you see any dangerous situations next to the track?
Yes   No

15. How would you rate your workload of the past 5 minutes on a scale?:

1  2  3  4  5  6  7
not         extremely
heavy       heavy
at all
Circle the right answer!

1. What signal are you approaching?
Green, green blinking + number, green blinking.
Yellow blinking + number, yellow number, yellow, yellow blinking, red

2. Is all the train equipment functioning properly?
Yes No

3. Are you giving traction?
Yes No

4. Are you fulfilling your breaking criterion?
Yes No

5. How would you rate your workload of the past 5 minutes on a scale?:
1 2 3 4 5 6 7
not extremely heavy heavy
at all

6. In comparison with the allowed speed, your speed is:
Too high Good

7. What type of brakes does this train have?
disc blocks magnetic electro-dynamic

Decision-making based on measurements of behavioural safety
8. What is the speed of the train driving in front of you?:

- Is driving slower than my train
- Is driving the same speed
- Is driving faster than my train
- There is no train driving in front of me

9. Did you see a signal out?

- Yes
- No

10. What is your train number?

- 2345
- 4321
- 1235
- 1234

11. What is your current speed? (Put an X at your speed.)

- 0
- 40
- 60
- 80
- 100
- 130
- 140

12. Does the current weather have an impact on the driving characteristics of the train?

- Yes
- No

13. Is de dead man’s switch functioning properly?

- Yes
- No

14. Are the cabin lights functioning?

- Yes
- No

15. Are the brakes working properly?

- Yes
- No
Decision-making based on measurements of behavioural safety

Driver:_________  Scenario:_________  Driver/Observer  Date ___/5/2013  Page 2B

Circle the right answer!

1. Are you driving your usual route?
   - Yes
   - No

2. What is the signal behind the signal you are approaching?
   - Green, green blinking + number, green blinking,
   - Yellow blinking + number, yellow number, yellow, yellow blinking, red

3. Are you braking?
   - Yes
   - No

4. Is the train delayed by more than three minutes?
   - Yes
   - No

5. Are the windshields broken?
   - Yes
   - No

6. Did you pass any defect level crossings?
   - Yes
   - No

7. How would you rate your alertness in the last five minutes on a scale?
   - 1 2 3 4 5 6 7
   - extremely
   - extremely sleepy
   - alert
8. Is the air conditioning functioning properly?
Yes  No

9. What do you expect that the coming signals will allow you?:
I can speed up  I can keep this speed  I will have to break

10. Did you see any obstacles on the track?
Yes  No

11. Did an ATB error occurred?
Yes  No

12. Did you see any construction near the track?
Yes  No

13. What was your previous signal?
Green,  green blinking +number,  green blinking,
Yellow blinking + number,  yellow number,  yellow,  yellow blinking,  red

14. How would you rate your workload of the past 5 minutes on a scale?:
1  2  3  4  5  6  7
not  extremely
heavy  heavy
at all

15. Is de loud-speaker installation functioning?
Yes  No


GROENEWEG, J. A. October 2013. RE: Personal communication.


behavioural safety on measurements of Decision


VAN DER HORST, A. & HOGEMA, J. 1994. Time-to-collision and collision avoidance systems. VERKEERSGEDRAG IN ONDERZOEK.


Measuring Situation Awareness of train drivers

R.J. Van der Meij BSc.