The identification of incidental learning as a cause of human error by exploring big data within railway safety

Burggraaf, J.M.

DOI
10.4233/uuid:c5b1a63b-3873-4b1b-b4ed-e12390d21d40

Publication date
2023

Document Version
Final published version

Citation (APA)
Burggraaf, J. M. (2023). *The identification of incidental learning as a cause of human error by exploring big data within railway safety*. [Dissertation (TU Delft), Delft University of Technology].
https://doi.org/10.4233/uuid:c5b1a63b-3873-4b1b-b4ed-e12390d21d40

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

The identification of incidental learning as a cause of human error by exploring big data within railway safety

Dissertation

for the purpose of obtaining the degree of doctor

at Delft University of Technology

by the authority of the Rector Magnificus prof. dr. ir. T.H.J.J. van der Hagen

chair of the Board of Doctorates

to be defended publicly on

Tuesday 17 January 2023 at 12:30 o'clock

by

Julia BURGGRAAF

Master of Science in Psychology, Leiden University, the Netherlands

Born in Beesel, the Netherlands

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus	Chairperson
Prof.dr.ir. P.H.A.J.M. van Gelder	Delft University of Technology, promotor
Prof.dr. J. Groeneweg	Delft University of Technology, promotor

Independent members:

Dr. G.P.H. Band	Leiden University, Netherlands
Prof.dr.ir. C. van Gulijk	University of Huddersfield, United Kingdom
Prof.dr. M.P. Hagenzieker	Delft University of Technology
Dr.ir. J. van den Top	Human Environment and Transport Inspectorate, Netherlands
Prof.dr.ir. G.L.L.M.E. Reniers	Delft University of Technology, reserve member

ProRail

“Many a false step was made by standing still.”

- Fortune cookie

Summary

Understanding human behavior and preventing accidents

Accidents at work come at a major cost including fatalities, disability and economic burden. The ideas around accident causation have changed over time from describing accidents as 'acts of gods' and as the fault of individual employees, to accidents being the result of an interaction between organizational, technical and human factors. Included in the more recent ideas is the notion that organizations have a role and responsibility in preventing accidents. Prevention can include eliminating error-promoting factors or adding safety barriers to prevent errors from leading to accidents.

When employees interact with any system in their organization, interventions can be aimed at employees and at the system. When investigating possibilities to improve the system, it is important to take into account how employees interact with the system. We need to be able to predict human behavior and in order to do that, we need to understand human behavior.

This book about my PhD research contributes to the expansion of knowledge on human behavior in two manners:

1. **New scientific knowledge on human behavior**
   The probability of human error increases when an employee is exposed to the same task circumstances in previous days followed by exposure to (visually) similar task circumstances that require different behavior.
2. **Insights into how big data* of actual human behavior (as captured by sensors) can be used in combination with psychological expertise to identify and answer other questions about human behavior in the future:**
   1. Analyzing the data with a focus on identifying discrepancies in the expected amount of variation in the behavior per factor versus the actual amount of variation leads to the identification of relevant research questions on human behavior 2. The task of using big data is itself a complex task which should also be considered from a Human Factors perspective** to decrease the chance that errors occur during the use of big data. Theory on cognitive biases and automatic activation can be used to identify pitfalls in complex tasks such as these. In this PhD-research, five pitfalls were identified to be aware of.

* Big data: data for which the quantity is too large to collect and process via traditional

** From a Human Factors perspective, human errors are not seen as an indication that there is something wrong with the individual who made the error but rather as an indication that parts of the system can be improved. Human Factors experts use the knowledge of human strengths and limitations, both mental and physical, to improve human-system interaction in such a way that safety, performance and/or employee satisfaction is enhanced. When trying to investigate why human errors occur, the first avenue is thus always to consider what factors within the system contributed to the causation of this error and how the system can be improved to prevent errors from occurring the future.

Opportunity for big data research on human behavior thanks to big data availability and questions within Dutch rail

A wide range of human behavior has been studied within social sciences and within industry, using different methodologies. Advances in technology make it possible to add a new methodology: analyzing the actual process related behavior of employees as captured by sensors that are already present or can be embedded within the task environment. This type of data is a type of big data, because the quantity is too large to collect and process via traditional technology.

Using big data within organizations has a lot of potential benefits, but these benefits are not automatically reaped. Turning the raw data into usable knowledge is not a straightforward process, especially since the data is often collected for different purposes than the research goal. Big data is also not commonly applied yet within industry to solve safety problems. It is possible to analyze big data of actual employee behavior without using expertise on human behavior to guide the analysis, but this does not necessarily lead to valuable insights with respect to human behavior that can actually be used to improve organizational processes and the accompanying level of safety via structural interventions.

Fortunately, the opportunity arose to use big data within the context of Dutch rail to study human behavior in a safety context. ProRail, the Dutch rail infrastructure manager, aims to reduce the amount of SPADs within the Netherlands. SPADs, or Signal Passed at Danger events, are incidents where a train passes a signal (with a red aspect) without authorization. Data was available on train driver deceleration behavior.

Five steps were taking that allowed us to identify a relevant research question and obtain valuable insights on human behavior that can be used in the prevention of accidents within rail and other industries.

Step 1. Analyzing the discrepancies between expected and actual level of variation in behavior to identify relevant (human) factors (chapter 5)

To be able to utilize the big data for safety purposes, we had to perform initial analyses in such a way that it would lead to insights that can be used for deeper analysis. Within psychology there has traditionally been a strong focus on analyzing differences in averages (mean or median). However, new causes of errors can be identified and their impact quantified by examining variation in system performance, rather than focusing on averages or only on problematic performance.

Examining variation includes asking questions such as: under which circumstances are there different amounts of variation in system performance (for example demonstrated as a larger standard deviation or wider curve), and is this difference in variation expected or not? Such questions are hard to answer with smaller sample sizes, especially when comparing the amount of variation across different conditions and testing very specific hypotheses.

Step 2. Decreasing the probability of errors during big data utilization by being aware of possible cognitive pitfalls in complex tasks such as data verification (chapter 6)

Human Factors specialists using big data to investigate human factors topics should be aware that they themselves are also susceptible to making errors. This raises the

questions: can 'we' (Human Factors experts) also use our knowledge of human factors to decrease the probability of errors during the use of big data?

One of the complex tasks within big data utilization is data verification. This step is especially important when using data for safety related purposes. Within the safety domain there is often a higher need for certainty due to high stakes and because it is often important to understand the causal mechanism to implement the correct intervention that does not have unwanted side-effects.

Using big data provides unique challenges in identifying data quality problems. It is not possible to verify every data point within big data, but it is possible to perform data verification by examining the data on certain aspects, including outliers, impossible combinations of data values and applying a four-eyes-principle on the methods used to process the data. Data verification is a task that can be affected by human error. In order to improve the data verification process, it is important to take cognitive biases into account.

In this thesis, five cognitive biases are listed that can occur during such data verification and thereby limit the identification of data quality problems. These biases manifest as pitfalls that are specified as 'The good form as evidence-error', 'The improved-thus-correct fallacy', 'Situation-dependent-identity-oversight', 'Impact underestimation' and 'The beaten track disadvantage' (section 6.3). The verification process can be improved with specific measures per pitfall to mitigate their effect and increase the probability of identifying data quality problems. These measures include incorporating specific checkpoints and questions within the verification process.

> It should be noted that there are standards within railways that ensure that safety critical software is made as safe as possible. The data that was used in this research was not part of any software with a safety critical component. There are also many different ways in which data verification can be performed. Chapter 6 gives insight into pitfalls that can be present during the data verification process of data that is available but does not yet comply to the strict standards applied that are relevant for safety critical software.

Other big data related tasks such as result visualization and interpretation can also be examined for the presence of common pitfalls. This is beyond the scope of this dissertation, but the principles applied in chapter 6 can be used to also inspect other tasks within big data utilization from a Human Factors perspective.

Step 3. Deciding which (human) factor to investigate further

Analyzing the data as described in step 1 pointed towards incidental learning as a potential cause of human error. Incidental learning is the non-intentional learning that automatically occurs during daily interaction with our surroundings.

The potential impact of incidental learning was chosen as an ideal candidate to study as part of the main research question of this dissertation because it is a factor:

- which has not been investigated yet in the context of human error
- which can be influenced by an organization in line with the Human Factors perspective of improving systems to reduce the probability of errors and increase safety, performance and employee satisfaction
- which is difficult to study using other methodologies
- which relates to fundamental human psychology and thus increased knowledge about this factor is not only beneficial for the railways but for all industries

Step 4. Investigating the impact of the factor on the behavior and remaining safety margins (chapter 2)

The impact of incidental learning was first analyzed using behavioral data as the dependent variable. This rich data source allowed the testing of nuanced hypotheses. More details on the analysis itself and the results are listed below in this summary in the section on incidental learning.

Big data can be analyzed using a 'bottom-up' approach where for example machine learning is used to find the most significant variables. We combined big data with a 'top-down' approach guided by content expertise, in this case human factors knowledge and theory about incidental learning, to:

- Improve the measure of human behavior
- support the data verification by identifying unlikely behavior patterns that could be caused by data quality problems
- determine which factors to include in analysis
- determine how the factors should be operationalized
- decide which additional factors to include in subsequent analyses (see step 5)

The calculation of multiple factors in the studies described in chapter 2 and 3 required such a specific combination of variables that it is unlikely to have been found by a big data analyst or team of analysist without thorough interaction with experts on human error (Section 7.1.4).

Step 5. Investigating the impact of the factor on incident occurrence (chapter 3)

The knowledge gained in step 4 was used to decide how to design the follow-up research where incident data was used as the dependent variable. This data source makes it possible to study whether an error also actually leads to accidents. Multiple factors can influence whether an error will indeed lead to an incident or could be corrected in time. To expand the knowledge of incidental learning of error and incident cause, we considered the role of self-correction. We calculated the 'opportunity for correction': the amount of room present for self-correction by the train driver as influenced by the infrastructure. More details on the analysis itself and the results are listed below in this summary in the section on incidental learning.

Summary of the discovered insights around incidental learning as a cause of human error

Incidental learning is the non-intentional learning that occurs automatically every day. A specific form of incidental learning is considered, namely the strengthened neural activation for certain behavior upon perceiving a cue or stimulus.

Whilst learning in itself is generally a positive term, it is hypothesized that this learning can lead to a 'usual response bias' which allows for more efficient responses in familiar situations, but can also lead to errors in specific situations. If there is indeed a significant negative influence of incidental learning, this is important to understand as it can undermine results of explicit training and awareness campaigns (See **Figure 1**). It is of course important to train employees (top right of **Figure 1**), but if incidental learning teaches employees different behavior (bottom left of **Figure 1**), then this explicit training will be partly undone.

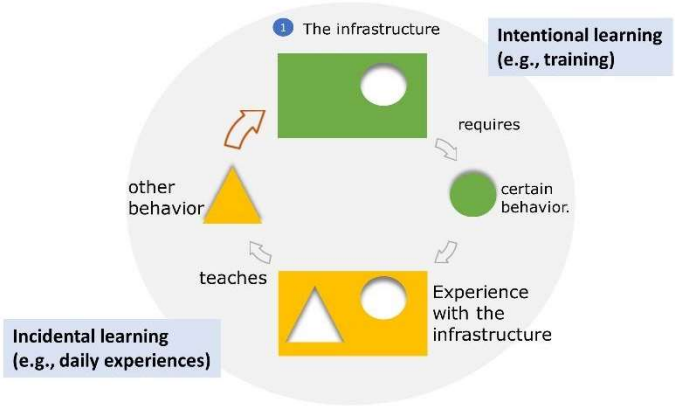


Figure 1. Differences in learning during intentional learning and incidental learning can lead to human error despite thorough explicit training. The green resembles an employee receiving explicit training on how he or she should perform according to company standards. The yellow resembles what an employee actually experiences on a day-to-day basis. If these differ, then human error can occur despite the employee successfully following the intentional learning sessions.

In the Netherlands, train drivers are exposed to multiple sources of variation in the Dutch infrastructure, including different distances between signals, but also different signal aspects. The same signal can for example show a green, yellow or red aspect, but also a yellow aspect with a number. The following research question was posited:

- To what extent and under which infrastructure related circumstances does incidental learning have a negative impact on train driver behavior during red aspect approaches?

A higher frequency of (signal aspect) exposure in combination with a high visual similarity is expected to increase the probability of an error occurring in situations where other deceleration behavior is required than the usual response. In Dutch rail, this situation can occur when a specific signal often shows a yellow+number aspect which requires relatively low amounts of deceleration (See blue rectangle in **Figure 2**), but the exact same signal sometimes also shows a 'plain' yellow aspect that requires a higher amount of deceleration (See red rectangle in **Figure 2**).

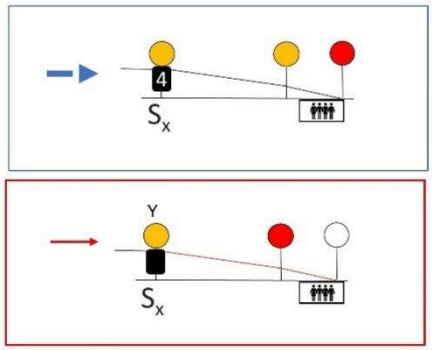


Figure 2. There can be variation in the signal aspect displayed by the same signal. In the approach at the top of the figure (in blue rectangle), the first signal has aspect yellow+4 indicating a speed reduction to 40 km/h because the signal at the station is red and the distance between the last two signals is insufficient for a green–yellow–red sequence. In the bottom approach (in red rectangle), the first signal has a yellow aspect because the next signal has a red aspect. The same amount of deceleration used after yellow+4 as visualized in the blue rectangle, is insufficient if used after yellow.

Data of actual train driver behavior in the Netherlands was studied. To study the effects of incidental learning, the train driver behavior was operationalized in two different manners for passenger trains. First, data of train speed and location during Red Aspect Approaches (RAAs) was used and transformed in a safety indicator that calculated the deceleration required to prevent a SPAD (chapter 2). Data on exposure to yellow+number aspects over the past fourteen days prior to the day of a RAA with green-yellow-red sequence was therefore used as a measure of incidental learning opportunity.

Subsequently, incident data was used, namely data on SPADs, in combination with data on the number of RAAs without SPADs (chapter 3). All of this data was already being recorded, although not immediately in useful formats. No additional sensors or other methods for data collection were required. The insights gained from the analysis of the behavior data were used to include an operationalization of *the train driver's opportunity to correct his or her initial error of insufficient deceleration* in the research design of the subsequent study. This factor was called "the window for correction" and could be operationalized by using the already available raw data in a different manner.

Various hypotheses with different sets of signals were tested using a simulation approach and a significance level of 0.05. The study using train driver deceleration behavior included sample sizes of 3429 RAAs, 1287 RAAs and a relatively small sample size of 415 RAAs when testing the hypothesis for one specific signal. The study using SPAD data included a sample of 1,139,665 RAAs and 29 relevant SPADs over a period of six years.

The results of both studies indicated a significant effect of previous yellow+number aspect exposure on train driver behavior and SPADs respectively, supporting the hypothesis that incidental learning can indeed have a negative impact on train driver behavior. The significant effects were substantial, including six times more SPADs per 100.000 RAAs for approaches with a high frequency of yellow+number aspects in the past fourteen days versus when there were no yellow+number aspects in the past fourteen days.

The study using incident data further showed that, in line with the hypothesis of incidental learning, the increase in incident probability only applies when the infrastructure characteristics are such that there is a smaller window for correction. There were 777,510 RAAs and 0 SPADs for the large window for correction, 319,533 RAAs and 3 SPADs for the medium window and 54,462 RAAs and 17 SPADs for the small window.

These results lead to the conclusion that the impact of incidental learning via previous exposure is an important factor to consider during accident analysis and, even more importantly, in system and task analysis. Current and future high-SPAD probability locations can be identified within Dutch rail if specific questions are answered affirmatively on possible aspect sequences and frequencies in specific locations, the size of the 'window for correction' as influenced by track speed, signal distance and aspects, and the presence of other SPAD prevention mechanisms.

Interventions to prevent SPADs can be aimed at preventing the initial error of insufficient deceleration due to incidental learning, increasing the opportunity for self-correction and/or implementing intervention mechanisms. Possible interventions include improved infrastructure design via adjusted signal placement and/or track speed, scheduling adjustments that impact which signal aspects are shown, adjusted signal aspect design, increased braking power, increased aspect visibility and technical intervention systems.

Whilst many insights were obtained about incidental learning as a contributor to accidents, there are still unanswered questions, mainly around the exact relation between exposure frequency and error probability, individual differences and around other types of contextual similarities than those investigated in this dissertation (See section 4.1.2.). A specific behavior can for example also be activated by visual similarities in other locations (e.g. different warehouse, but similar tools and working environment) or auditory similarities (e.g. similarly sounding alarms).

These questions can be answered in future research to provide new insight into human behavior and concrete pointers for the rail industry and other industries on how organizations can improve their processes to support their employees and reduce the probability of errors and accidents.

Figure 3 displays an oversimplified explanation of incidental learning to help the reader who is not yet familiar with the topic by proving an overview.

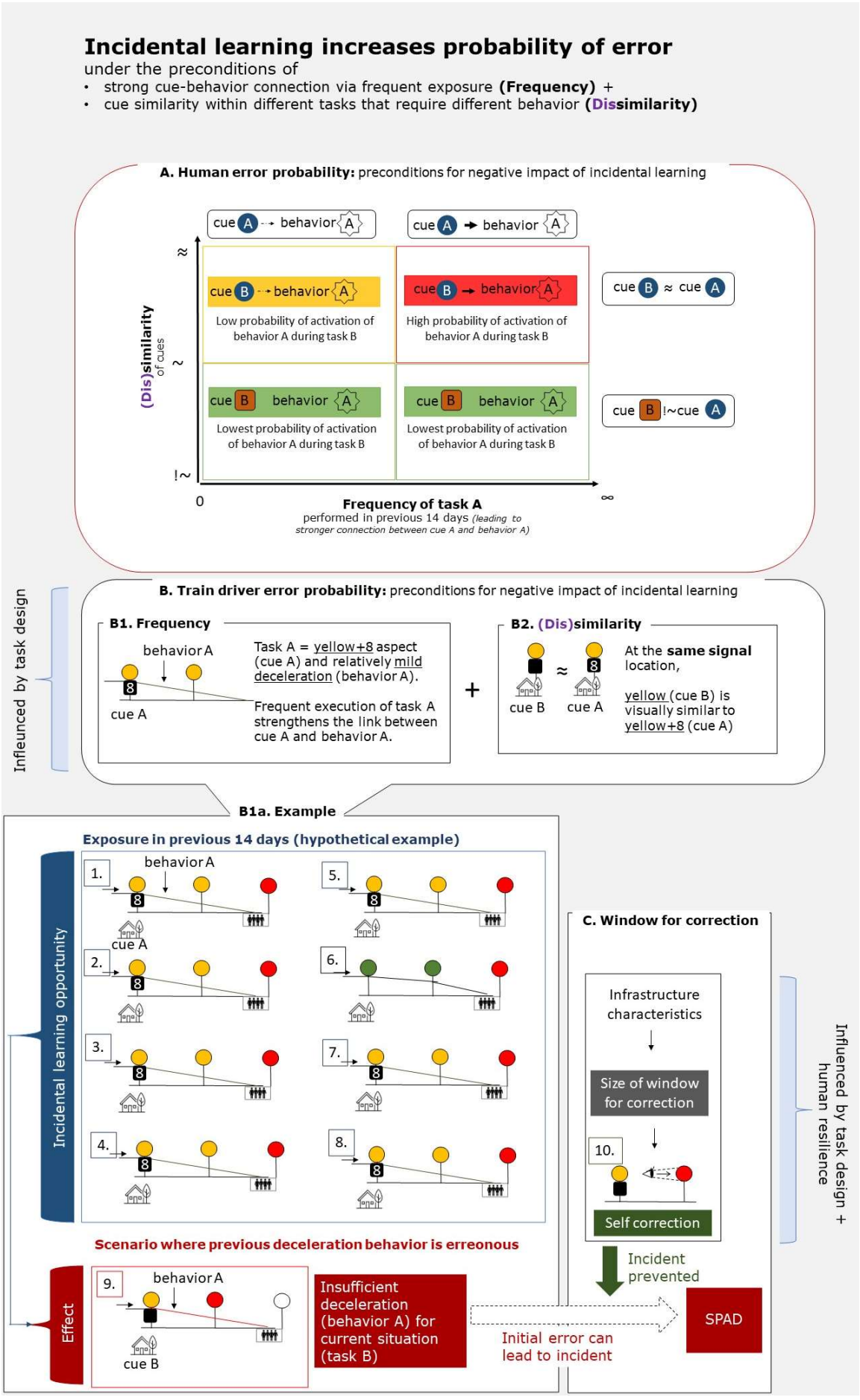

Incidental learning increases probability of error
under the preconditions of
• strong cue-behavior connection via frequent exposure (Frequency) +
• cue similarity within different tasks that require different behavior (Dissimilarity)
A. Human error probability: preconditions for negative impact of incidental learning
cue A behavior A
cue A behavior A
cue B behavior A
Low probability of activation of behavior A during task B
cue B behavior A
High probability of activation of behavior A during task B
cue B ≈ cue A
cue B behavior A
Lowest probability of activation of behavior A during task B
cue B behavior A
Lowest probability of activation of behavior A during task B
cue B ≠~ cue A
(Dis)similarity of cues
≈
~
≠~
0
∞
Frequency of task A
performed in previous 14 days (leading to stronger connection between cue A and behavior A)
Influenced by task design
B. Train driver error probability: preconditions for negative impact of incidental learning
B1. Frequency
behavior A
cue A
Task A = yellow+8 aspect (cue A) and relatively mild deceleration (behavior A).
Frequent execution of task A strengthens the link between cue A and behavior A.
+
B2. (Dis)similarity
cue B
≈
cue A
At the same signal location,
yellow (cue B) is visually similar to yellow+8 (cue A)
B1a. Example
Exposure in previous 14 days (hypothetical example)
Incidental learning opportunity
behavior A
cue A
1.
2.
3.
4.
5.
6.
7.
8.
Scenario where previous deceleration behavior is erreonous
Effect
9.
behavior A
cue B
Insufficient deceleration (behavior A) for current situation (task B)
Initial error can lead to incident
C. Window for correction
Infrastructure characteristics
Size of window for correction
10.
Self correction
Incident prevented
SPAD
Influenced by task design + human resilience

Figure 3. Visual summary of incidental learning as error cause.

Section A illustrates under which preconditions incidental learning can lead to an error. Consider that correct execution of 'task A' consists of perception of 'cue A', followed by performance of 'behavior A', and correct execution of 'task B', consists of perception of 'cue B' followed by performance of 'behavior B'. Incidental learning can lead to an error during task B when 'behavior A' is activated instead of 'behavior B' after perception of 'cue B'.

This can occur IF

- 'cue B' and 'cue A' are similar ánd
- perception of 'cue A' strongly activates 'behavior A' as a result of frequent exposure to and execution of 'task A'

Section B shows one example of how these preconditions can take place in Dutch rail, using the example of signal aspect 'yellow+8'.

The signal aspect yellow+8 is 'cue A', with a mild deceleration as 'behavior A'. The signal aspect yellow is cue B with 'behavior B' being a more stronger deceleration in order to stop in front of the red aspect.

Section B2 visualizes the similarity between signal aspect yellow+8 and yellow at the same signal location. **Section B1** visualizes the frequency of task A which influences the strength with which perception of yellow+8 (cue A) activates the mild deceleration (behavior A).

Section B1a gives a simplified example of a train driver being exposed to and executing task A seven times (approaches 1-5 and 7-8). During approach 9, the yellow aspect (cue B) is present as part of task B with the correct behavior being a stronger deceleration (behavior B), but mild deceleration (behavior A) is performed which is an error in this setting.

Section C. shows that the error of insufficient deceleration can either lead to a SPAD or not. A SPAD can be prevented if the train driver self-corrects upon perceiving the red aspect. Timely self-correction is only likely IF there is sufficient opportunity for self-correction. The size of the window for correction is influenced by the infrastructure design (i.e. the task design).

The two blue brackets to the side of section B and C highlight that both the occurrence of initial insufficient deceleration and self-correction are influenced by infrastructure design (i.e. the task design).

The insights gained around incidental learning show that the use of big data in combination with Human Factors expertise can lead to valuable new insights

The application of big data led to a gain in knowledge on incidental learning, specifically that it can indeed negatively impact employee behavior and that this effect occurs under specific circumstances that are influenced by the work environment as created by the organization.

Thanks to the availability and use of big data of actual employee behavior, research on incidental learning can be expanded to behavior and to considering the potential negative impact of incidental learning. The field of Human Factors can be expanded by considering not only whether a task can be performed at a given moment, but also by including previous exposure in that evaluation. It is recommended to consider a potential role of incidental learning during accident analysis and during task evaluation and task design (See section 4.2).

The insights gained thanks to big data research within Dutch rail show that the use of big data of actual employee behavior can indeed enrich our understanding of human behavior in new and more detailed ways. By using this type of research, we can expand the scientific body of knowledge concerning safety, whilst simultaneously crossing the bridge from academia towards industry. This increases the chance of implementation and eventual improvements in safety of those performing the work day in, day out.

Samenvatting

Menselijk gedrag begrijpen en ongevallen voorkomen

Ongevallen op het werk kunnen aanzienlijke gevolgen hebben waaronder dodelijke slachtoffers, verwondingen en economische lasten. De ideeën over het ontstaan van ongevallen zijn in de loop der tijd veranderd: ongevallen worden niet langer omschreven als 'acts of god' en als de schuld van de individuele werknemer, maar als het resultaat van een wisselwerking tussen organisatorische, technische en menselijke factoren. Daarbij wordt de rol en verantwoordelijkheid van organisaties in het voorkomen van ongevallen erkend. Interventies kunnen bestaan uit het elimineren van factoren die de kans op fouten verhogen of uit het toevoegen van veiligheidsbarrières om te voorkomen dat fouten tot ongevallen leiden.

Aangezien er tijdens werkprocessen een interactie is tussen werknemers en het systeem, kunnen interventies gericht worden op werknemers en op het systeem. Bij het onderzoeken van mogelijkheden om het systeem te verbeteren is het belangrijk om rekening te houden met de manier waarop werknemers en het systeem elkaar wederzijds beïnvloeden. We moeten daarvoor menselijk gedrag begrijpen.

Dit boek over mijn promotieonderzoek draagt op twee manieren bij aan de uitbreiding van kennis over menselijk gedrag:

1. **Nieuwe wetenschappelijke kennis over menselijk gedrag**
   De kans op menselijke fouten neemt toe wanneer een werknemer in voorgaande dagen is blootgesteld aan dezelfde taakomstandigheden, gevolgd door blootstelling aan (visueel) vergelijkbare taakomstandigheden die ander gedrag vereisen.
2. **Inzichten in hoe big data* van daadwerkelijk menselijk gedrag (zoals gemeten door sensoren) kunnen worden gebruikt in combinatie met psychologische theorie om in de toekomst andere vragen over menselijk gedrag te identificeren en te beantwoorden:**
   1. Relevante onderzoeksvragen over menselijk gedrag kunnen geïdentificeerd worden door de data te analyseren met een focus op discrepanties in de verwachte hoeveelheid variatie in gedrag per factor versus de werkelijke hoeveelheid variatie in gedrag voor die factor. 2. Het gebruik van big data is een complexe taak op zichzelf die ook vanuit een Human Factors perspectief ** bekeken kan worden om de kans te verkleinen dat er fouten optreden tijdens het gebruik van big data. Theorie over denkfouten en automatische activering kan worden gebruikt om valkuilen in complexe taken als deze te identificeren. In dit promotieonderzoek zijn vijf valkuilen geïdentificeerd waarop met moet letten tijdens de verificatie van big data.

* Data van te grote kwantiteit om via traditionele technologie te verzamelen en te verwerken.

** Vanuit het Human Factors perspectief worden menselijke fouten niet gezien als een aanwijzing dat er iets mis is met het individu dat de fout heeft gemaakt, maar eerder als aanwijzing dat delen van het systeem kunnen worden verbeterd. Human Factors experts gebruiken de kennis van de sterke en zwakke punten van de mens, zowel mentaal als fysiek, om de interactie mens en systeem zodanig te verbeteren dat de veiligheid, de prestaties en/of de tevredenheid van de werknemers worden verbeterd. Wanneer men probeert te onderzoeken waarom menselijke fouten optreden, moet men dus altijd eerst nagaan welke factoren binnen het systeem hebben bijgedragen tot het ontstaan van deze fout en hoe het systeem kan worden verbeterd om fouten in de toekomst te voorkomen.

Kansen voor onderzoek naar menselijk gedrag met big data dankzij big data beschikbaarheid en vraagstukken binnen het Nederlandse spoor

Menselijk gedrag wordt binnen de sociale wetenschappen en binnen de industrie onderzocht aan de hand van verschillende methodes. Vooruitgang in de technologie maakt het mogelijk om een nieuwe methode toe te voegen: het analyseren van daadwerkelijk procesgerelateerd gedrag van werknemers dat gemeten wordt door sensoren die al aanwezig zijn of kunnen worden ingebed in de taakomgeving. Dit soort data is een vorm van big data, omdat de kwantiteit te groot is om via traditionele technieken te verzamelen en te verwerken.

Het gebruiken van big data binnen organisaties om meer te leren over menselijke gedrag heeft veel potentiële voordelen, maar deze voordelen worden niet automatisch benut. Het omzetten van de ruwe data in bruikbare kennis is geen rechttoe-rechtaan proces, vooral omdat de data vaak voor andere doeleindes dan het onderzoeksdoel wordt verzameld. Big data wordt momenteel ook nog niet vaak ingezet door de industrie om hun veiligheidsvraagstukken op te lossen. Het is mogelijk om big data van daadwerkelijk werknemersgedrag te analyseren zonder psychologische theorie te gebruiken om de analyse te sturen, maar dit leidt niet noodzakelijkerwijs tot waardevolle inzichten die ook daadwerkelijk kunnen worden gebruikt om organisatieprocessen en het bijbehorende veiligheidsniveau te verbeteren via structurele interventies.

Er deed zich gelukkig de gelegenheid voor om big data binnen de context van het Nederlandse spoor te gebruiken om menselijk gedrag in een veiligheidscontext te bestuderen. ProRail, de Nederlandse spoorinfrastructuurbeheerder, wil het aantal STS-passages binnen Nederland verminderen. STS-passages, of StopTonend Sein passages, zijn incidenten waarbij een trein zonder toestemming een sein (met een rood seinbeeld) passeert. Data was beschikbaar over het remgedrag van machinisten richting stoptonende seinen.

In vijf stappen konden we een relevante onderzoeksvraag identificeren en waardevolle inzichten verkrijgen over menselijk gedrag die gebruikt kunnen worden bij de preventie van ongevallen binnen de spoorwegen en andere industrieën.

Stap 1. Analyseren van de discrepanties tussen verwacht en werkelijk niveau van variatie in gedrag om relevante (menselijke) factoren te identificeren

Om big data voor veiligheidsdoeleinden te kunnen gebruiken, moesten we de eerste analyses zodanig uitvoeren dat ze zouden leiden tot inzichten die gebruikt kunnen worden voor diepere analyses. Binnen de psychologie is er traditioneel een sterke focus op het analyseren van verschillen in centrummaten (gemiddelde of mediaan). Nieuwe oorzaken van fouten kunnen echter worden opgespoord en hun impact gekwantificeerd door de variatie in systeemprestaties te onderzoeken, in plaats van zich te richten op gemiddelden of alleen op problematische prestaties.

Het onderzoeken van variatie omvat vragen als: onder welke omstandigheden zijn er verschillende hoeveelheden variatie in de systeemprestaties (bijvoorbeeld aangetoond als een grotere standaardafwijking of bredere curve), en is dit verschil in variatie te verwachten op basis van hoe het systeem beoogt te werken? Dergelijke vragen zijn moeilijk te beantwoorden met kleinere steekproefgrootten, vooral wanneer de hoeveelheid variatie in verschillende omstandigheden wordt vergeleken en zeer specifieke hypothesen worden getest.

Stap 2. De kans op fouten bij het gebruik van big data verkleinen door zich bewust te zijn van mogelijke cognitieve valkuilen bij complexe taken zoals dataverificatie

Human Factors experts die big data gebruiken om human factors onderwerpen te onderzoeken moeten zich ervan bewust zijn dat zij zelf ook vatbaar zijn voor het maken van fouten. Dit roept de vraag op: kunnen 'wij' (Human Factors experts) onze kennis van menselijke factoren ook gebruiken om de kans op fouten bij het gebruik van big data te verkleinen?

Een van de complexe taken binnen het gebruik van big data is de verificatie van data. Deze stap is vooral belangrijk bij het gebruik van data voor veiligheidsgerelateerde doeleinden. Binnen het veiligheidsdomein is er vaak een grotere behoefte aan zekerheid vanwege de hoge inzet en omdat het vaak belangrijk is het causale mechanisme te begrijpen om de juiste interventie uit te voeren die geen ongewenste neveneffecten heeft.

Het gebruik van big data zorgt voor unieke uitdagingen bij het opsporen van problemen met de datakwaliteit. Het is niet mogelijk om elk datapunt binnen big data te verifiëren, maar het is wel mogelijk om data te verifiëren door de data te onderzoeken op bepaalde aspecten, waaronder uitschieters, onmogelijke combinaties van waarden en het toepassen van een vier-ogen-principe op de methoden die zijn gebruikt om de data te verwerken. Dataverificatie is een taak die door menselijke fouten kan worden beïnvloed. Om het dataverificatieproces te verbeteren is het belangrijk rekening te houden met denkfouten.

In dit proefschrift worden vijf denkfouten opgesomd die kunnen optreden tijdens een dergelijke dataverificatie en daardoor de identificatie van problemen met de datakwaliteit beperken. Deze denkfouten manifesteren als valkuilen genaamd ‘De goede vorm als bewijs-fout’, ‘De verbeterd-dus-goed misvatting’, ‘Situatie-afhankelijke-identiteit-onzorgvuldigheid’, ‘Impact onderschatting’ en ‘Het gebaande-paden-nadeel’. Het verificatieproces kan verbeterd worden door specifieke maatregelen per valkuil toe te passen en zo de kans te vergroten dat aanwezige datakwaliteitsproblemen daadwerkelijk worden geïdentificeerd. Deze maatregelen bestaan uit het inbouwen van specifieke triggers en werkwijzen in de uitvoer van de verificatie.

Kanttekening: Binnen de spoorwegen bestaan normen die ervoor zorgen dat veiligheidskritische software zo veilig mogelijk wordt gemaakt. De data die in dit promotieonderzoek zijn gebruikt, maakten geen deel uit van software met een veiligheidskritisch component. Hoofdstuk 6 geeft inzicht in valkuilen die aanwezig kunnen zijn tijdens het dataverificatieproces van data die wel beschikbaar is, maar nog niet voldoet aan de strenge normen die gehanteerd worden die relevant zijn voor veiligheidskritische software. Daarnaast zijn er ook veel verschillende manieren waarop dataverificatie kan worden uitgevoerd. De beschreven manieren in hoofdstuk 6 zijn niet bedoeld als uitputtende lijst.

Ook andere big data gerelateerde taken, zoals visualisatie en interpretatie van resultaten, kunnen worden onderzocht op de aanwezigheid van veelvoorkomende valkuilen. Dit valt buiten het bestek van dit proefschrift, maar de in hoofdstuk 6 toegepaste principes kunnen worden gebruikt om ook andere taken binnen big data-gebruik vanuit een Human Factors-perspectief te inspecteren.

Stap 3. Beslissen welke (menselijke) factor verder onderzocht moet worden

Het analyseren van de data zoals beschreven in stap 1 wees in de richting van incidenteel leren als mogelijke oorzaak van menselijke fouten. Incidenteel leren is het niet-intentionele leren dat automatisch plaatsvindt tijdens de dagelijkse interactie met onze omgeving.

De potentiële impact van incidenteel leren werd gekozen als ideale kandidaat om te bestuderen als hoofdonderzoeksvraag van dit proefschrift omdat het een factor is

- die nog niet is onderzocht in de context van menselijke fouten
- die door een organisatie kan worden beïnvloed in overeenstemming met het Human Factors perspectief van het verbeteren van systemen om de kans op fouten te verminderen en de veiligheid, prestaties en werknemerstevredenheid te verhogen
- die moeilijk te bestuderen via andere methoden
- die verband houdt met de fundamentele menselijke psychologie en dus is meer kennis over deze factor niet alleen gunstig voor de spoorwegen maar voor alle bedrijfstakken

Stap 4. Het onderzoeken van de impact van de factor op het gedrag en op de resterende veiligheidsmarges

Het effect van incidenteel leren werd eerst geanalyseerd met behulp van gedragsdata als afhankelijke variabele. Deze rijke databron maakte het mogelijk genuanceerde hypothesen te testen. Meer details over de analyse zelf en de resultaten staan hieronder in de samenvatting over incidenteel leren.

Big data kan worden geanalyseerd met een 'bottom-up' benadering waarbij bijvoorbeeld machine learning wordt gebruikt om de meest significante variabelen te vinden. Wij combineerden big data met een 'top-down' benadering geleid door inhoudelijke expertise, in dit geval human factors kennis en theorie over incidenteel leren, om:

- de meting van menselijk gedrag te verbeteren
- de dataverificatie te ondersteunen door onwaarschijnlijke gedragspatronen te identificeren die zouden kunnen worden veroorzaakt door problemen met de datakwaliteit
- te bepalen welke factoren in de analyse moeten worden opgenomen
- te bepalen hoe de factoren moeten worden geoperationaliseerd
- beslissen welke aanvullende factoren in latere analyses moeten worden opgenomen (zie stap 5)

De berekening van meerdere factoren in de in hoofdstuk 2 en 3 beschreven studies vereiste zo'n specifieke combinatie van variabelen dat het onwaarschijnlijk is dat deze door een big data-analist of een team van analisten zou zijn gevonden zonder grondige interactie met deskundigen op het gebied van menselijke fouten.

Stap 5. Onderzoek naar het effect van de factor op het optreden van incidenten

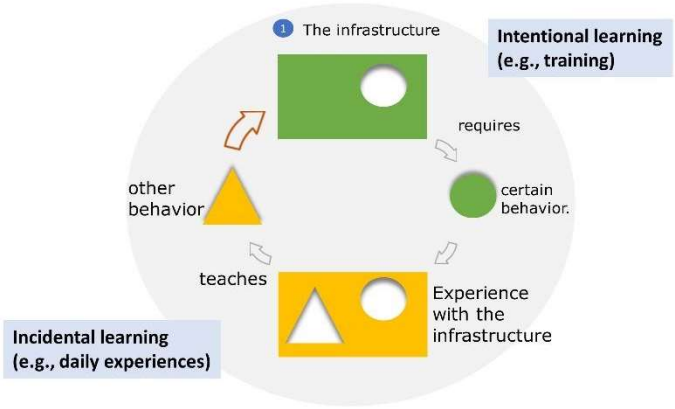
De in stap 4 opgedane kennis is gebruikt om te beslissen hoe het vervolgonderzoek moet worden opgezet waarbij incidentdata als afhankelijke variabele worden gebruikt. Deze databron maakt het mogelijk te bestuderen of een fout ook daadwerkelijk tot ongevallen leidt. Meerdere factoren kunnen van invloed zijn of een fout inderdaad tot een incident

leidt of tijdig kan worden gecorrigeerd. Om de kennis van incidenteel leren als fout- en incidentoorzaak uit te breiden, hebben we gekeken naar de rol van zelfcorrectie. We berekenden de 'gelegenheid tot correctie': de hoeveelheid ruimte die aanwezig is voor zelfcorrectie door de machinist. Deze aanwezige ruimte wordt beïnvloed door de infrastructuur. Meer details over de analyse zelf en de resultaten staan hieronder in de samenvatting over incidenteel leren.

Samenvatting van de ontdekte inzichten rond incidenteel leren als oorzaak van menselijke fouten

Incidenteel leren is het niet-intentioneel leren dat iedere dag automatisch plaatsvindt. Een specifieke vorm van incidenteel leren is bekeken, namelijk verhoogde neurale activatie voor bepaald gedrag bij het waarnemen van een cue of stimulus.

Hoewel de term 'leren' doorgaans wordt gezien als een positieve term, wordt in deze dissertatie de hypothese gesteld dat leren ook kan leiden tot een 'standaard reactie bias' welke het mogelijk maakt om efficiënter te reageren, maar ook kan leiden tot het maken van fouten in specifieke situaties. Als incidenteel leren inderdaad een significante negatieve invloed kan hebben, dan is het belangrijk om hiervan op de hoogte te zijn omdat het de resultaten van expliciete trainingen en bewustwordingscampagnes kan ondermijnen (zie **Figuur 1**). Het is natuurlijk belangrijk om werknemers te trainen (rechtsboven **Figuur 1**), maar als incidenteel leren werknemers ander gedrag aanleert (linksonder **Figuur 1**), dan zal de expliciete training gedeeltelijk teniet gedaan worden.



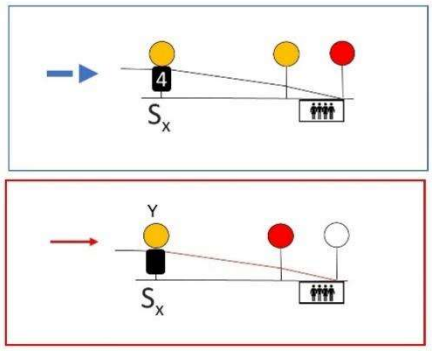
Figuur 1. Verschillen in het geleerde tijdens intentioneel en incidenteel leren kan leiden tot menselijke fouten ondanks grondige expliciete training. Het groen staat voor de expliciete training die een werknemer ontvangt over hoe hij of zij een taak moet uitvoeren. Het geel staat voor daadwerkelijke ervaring van de werknemer met de taak op een dagelijkse basis. Als deze twee verschillen, dan kan een fout optreden ondanks dat de werknemer de intentionele training succesvol heeft afgerond.

In Nederland worden machinisten blootgesteld aan verschillende bronnen van variatie in de infrastructuur, waaronder verschillende afstanden tussen seinen maar ook verschillende seinbeelden. Hetzelfde sein kan bijvoorbeeld een groen, geel of rood seinbeeld tonen, maar ook een geel seinbeeld met een getal.

De volgende onderzoeksvraag is gesteld:

- In welke mate en onder welke infrastructuur gerelateerde omstandigheden heeft incidenteel leren een negatieve impact op machinistenrijgedrag tijdens een roodseinnadering?

De verwachting is dat een hogere frequentie van (seinbeeld)blootstelling in combinatie met een grote visuele vergelijkbaarheid de kans verhoogt dat een fout wordt gemaakt wanneer ander remgedrag is vereist dan de standaard reactie. Binnen het Nederlandse spoor kan deze situatie optreden wanneer een specifiek sein vaak een geel+getal seinbeeld toont waarbij een relatief lage remvertraging voldoende is (zie blauwe rechthoek bovenin **Figuur 2**), maar waar soms ook een 'gewoon' geel seinbeeld wordt getoond waarbij een hogere remvertraging nodig is (zie rode rechthoek onderin **Figuur 2**).



Figuur 2. Er kan variatie zijn in het seinbeeld dat getoond wordt door hetzelfde sein. Tijdens de bovenste nadering heeft het eerste sein het seinbeeld geel+4 wat aangeeft dat een snelheidsvermindering naar 40 km/h uitgevoerd moet worden. Het seinbeeld is in dit geval geel+4 omdat het sein bij de halte een rood seinbeeld heeft en de afstand tussen de laatste twee seinen onvoldoende is voor een groen-geel-rood seinbeeldenreeks. Tijdens de onderste nadering toont het eerste sein het seinbeeld geel omdat het daaropvolgende sein rood toont. De hoeveelheid remvertraging die voldoende is tijdens de bovenste nadering, is onvoldoende tijdens de onderste nadering.

Data van daadwerkelijk machinistenrijgedrag in Nederland is geanalyseerd. Om de effecten van incidenteel leren te onderzoeken is het rijgedrag van reizigersmachinisten op twee verschillende manieren geoperationaliseerd. Eerst werden data over treinsnelheid en -locatie tijdens rood seinnaderingen (RSNs) gebruikt en omgezet in een veiligheidsindicator die de vertraging berekent die nodig is om een STS-passage te voorkomen (hoofdstuk 2). Data over de blootstelling aan geel+getal seinbeelden in de veertien dagen voorafgaand aan de dag van de RSN met een groen-geel-rood seinbeeldenreeks is daarom gebruikt als maat van gelegenheid tot incidenteel leren.

Vervolgens werd incidentdata gebruikt, namelijk data over STS-passages, in combinatie met data over het aantal RSNs zonder STS-passage (hoofdstuk 3). Al deze data werden reeds geregistreerd, zij het niet onmiddellijk in bruikbare formaten. Er waren geen extra sensoren of andere methoden voor dataverzameling nodig. De inzichten die verkregen waren uit de analyse van de gedragsdata werden gebruikt om een operationalisering toe te voegen van de ruimte die aanwezig was voor de machinist om diens initiële fout van onvoldoende remming te corrigeren. Deze factor wordt aangeduid als 'correctieruimte' en kon berekend worden door de reeds beschikbare ruwe data op een andere manier te gebruiken.

Het testen van verschillende hypotheses met verschillende selecties van seinen is gedaan aan de hand van een simulatieaanpak en een significantieniveau van 0.05. De studie met machinistenremgedrag bevatte samples van 3429 RSNs, 1287 RSNs en een relatief klein sample van 415 RSNs voor het testen van de hypothese voor één specifieke sein. In de studie met STS-passage data is data onderzocht van een periode van 6 jaar met daarin 1.139.665 RSNs en 29 relevante STS-passages.

In beide studies is een significant effect gemeten van eerdere geel+getal seinbeeldblootstelling op machinistenrijgedrag en de kans op een STS-passage. Deze bevindingen ondersteunen de hypothese dat incidenteel leren inderdaad een negatieve impact kan hebben op machinisten rijgedrag. De significante effecten waren substantieel, waaronder zes keer meer STS-passages per 100.000 RSNs voor naderingen met een hoge frequentie van geel+getal seinbeelden in de voorafgaande veertien dagen versus wanneer er geen geel+getal seinbeelden waren in de voorgaande veertien dagen.

De studie aan de hand van incidentdata liet verder zien dat, in lijn met de hypothese over incidenteel leren, de toename in kans op een STS-passage alleen aanwezig is wanneer de infrastructuurinrichting dusdanig is dat er weinig correctieruimte is. Er waren 777.510 RSNs en 0 STS-passages bij een grote correctieruimte, 319.533 RSNs en 3 STS-passages bij een medium correctieruimte en 54.462 RSNs en 17 STS-passages bij een kleine correctieruimte.

Deze resultaten leiden tot de conclusie dat de impact van incidenteel leren door seinbeeldblootstelling in voorafgaande dagen een belangrijke factor is om mee te nemen tijdens ongevalsonderzoek en in systeem- en taakanalyse. Huidige en toekomstige locaties met een hoge kans op STS-passages kunnen geïdentificeerd worden door een set aan vragen te beantwoorden waaronder vragen over mogelijke seinbeelden en seinbeeldfrequenties op die locatie, de grootte van de correctieruimte zoals beïnvloed door baanvaksnelheid, seinafstand en seinbeelden, en de aanwezigheid van andere STS-passage interventiemiddelen.

Interventies om STS-passages te voorkomen kunnen toegepast worden om de initiële fout van onvoldoende remming door incidenteel leren te voorkomen, verhoogde mogelijkheid te bieden voor zelfcorrectie door de machinist en/of het toepassen van andere corrigerende technieken. Mogelijke interventies omvatten aanpassing van infrastructuurontwerp via aangepaste seinplaatsing en/of baanvaksnelheid, dienstregelingsaanpassingen die beïnvloeden welke seinbeelden worden getoond, aangepast seinbeeldontwerp, toename in remvermogen, toename in seinzichtbaarheid en technische interventiesystemen.

Er zijn nog steeds open vragen rondom incidenteel leren als bijdragende factor aan ongevallen, namelijk rondom de exacte relatie tussen blootstellingfrequentie en foutkans, individuele verschillen en rondom andere soorten contextuele overeenkomsten dan degene onderzocht in deze dissertatie. Een bepaalde gedraging kan bijvoorbeeld ook geactiveerd worden door visuele overeenkomsten op andere locaties (bijv. ander magazijn, maar vergelijkbaar gereedschap en werkomgeving) of auditieve overeenkomsten (bijv. vergelijkbaar klinkende alarmen).

Het beantwoorden van deze vragen in toekomstig onderzoek zal waardevolle inzichten geven in menselijk gedrag en nog meer concrete handvatten bieden voor de spoorsector en andere sectoren over hoe organisaties hun processen beter kunnen inrichten zodat werknemers worden ondersteund en de kans op fouten en ongevallen wordt verkleind.

De verkregen inzichten rond incidenteel leren laten zien dat het gebruik van big data in combinatie met psychologietheorie tot waardevolle nieuwe inzichten kan leiden

De toepassing van big data leidde tot meer kennis over incidenteel leren, met name dat het inderdaad een negatief effect kan hebben op het gedrag van werknemers en dat dit

effect optreedt onder specifieke omstandigheden die worden beïnvloed door de werkomgeving zoals die door de organisatie wordt gecreëerd.

Dankzij de beschikbaarheid en het gebruik van data van daadwerkelijk gedrag kan onderzoek naar incidenteel leren uitgebreid worden naar gedrag en uitgebreid worden naar een negatieve impact van incidenteel leren. Het vakgebied Human Factors kan ook uitgebreid worden door niet alleen te beoordelen of een taak uitgevoerd kan worden op een specifiek moment, maar door ook blootstelling in voorafgaande dagen mee te nemen in de evaluatie van het taakontwerp. Het wordt aanbevolen om de rol van incidenteel leren ook mee te nemen tijdens ongevalsonderzoek en tijdens het ontwerp van taken en processen.

De inzichten die zijn opgedaan dankzij het gebruik van big data in het Nederlandse spoor laten zien dat het gebruiken van data van daadwerkelijk werknemersgedrag inderdaad onze kennis over menselijk gedrag kan verrijken op nieuwe en meer gedetailleerde wijze. Door dit soort onderzoek toe te passen, kunnen we de wetenschappelijke literatuur rondom veiligheid uitbreiden en tegelijkertijd een brug bouwen van de academische wereld richting de industrie. Dit verhoogt de kans dat de opgedane inzichten daadwerkelijk tot implementatie leiden en vervolgens een verbetering van de veiligheid van zij die het werk uitvoeren, dag in, dag uit.

Preface

Ever since I was a young, I have been tutoring, explaining and presenting. In primary school, I used to finish my schoolwork fairly quickly and would spend the remaining time helping classmates or daydreaming, depending on what was allowed. In high school, I earned some pocket money tutoring in math, physics and chemistry. In university, I earned some more pocket money tutoring my fellow psychology students in statistics and eventually I was hired at the university as a statistics teacher during my masters.

What I learned from these experiences is that I love figuring out a way to explain very complicated information in a more digestible manner and that by doing so, my own knowledge and understanding increased tenfold, leading to new insights. I also noticed that psychology students do not tend to love statistics, but I did. I even wrote my bachelor thesis on a statistics subject and went to one of those information meetings about the statistics master within psychology. Eventually I chose cognitive psychology as a master, because I realized that I mainly love data as a tool to tell me something about a psychological phenomenon and that part of my love for statistics came from being good at it, which I did not think was a good basis for a great love affair.

Two years after my graduation, I switched from consultancy jobs to an external PhD at the Technical University of Delft. Here I got the opportunity to combine two things I love: psychology and data. I also got to combine working within industry and within academia which allowed me to add my third love into the mix: figuring out a way to explain very complicated information in a more digestible manner. And finally, there was the fourth ingredient that I did not know I was missing until I spend two years working full time: being given the time and space to thoroughly investigate a complicated problem.

In the early phases of my PhD, I had to give a presentation which I thought went pretty well. My academic audience of two did not think so. The feedback I got was: 'you are not a data expert.' What the person was referring to, was that I am a cognitive psychologist by trade and that this did not become apparent in that particular presentation. In hindsight, I am grateful for that feedback as it forced me to refocus on writing from my main strength and passion, without excluding the other elements.

I have learned a lot from both the scientific community and the industry, including that there are many (sometimes frustrating) differences between the two but that they can also complement each other beautifully. This research could not have taken place without either. I have also tried to write this dissertation for both audiences.

To add value to both audiences, I have included both the peer reviewed scientific papers as well as additional sections. These additional sections might not be typical for a dissertation because they do not include all the nuances and careful wording that are included in the scientific papers. However they do add value to industry by including practical lessons learned and images that help to illustrate the message.[1]

All of this has led to the book before you. A book that has been written by a social scientist with characteristics of a data analyst, whilst spending most days within industry and pursuing answers as one does within academia. I have found tremendous value in the coming together of these worlds, as I hope you will too.

[1] The sections where I have taken more language and illustrative liberties to improve readability and accessibility for industry colleagues and encourage application of the research: parts of the main summary, the visual summaries shown in Figure 3, Figure 6, Figure 19 and chapter 7.

Contents

Chapter 1.

General introduction: Investigating human error with big data

1.1 This dissertation contributes to the aim of increasing the scientific knowledge of human behavior via two paths

I performed my PhD-research within the context of rail, but this dissertation is first and foremost about human behavior. My goal was to advance the (scientific) knowledge about human behavior and help prevent organizational accidents, irrespective of the industry. My PhD-research contributed to that goal via two main gains:

1. **New scientific knowledge on human behavior**
   The probability of human error increases when an employee is exposed to the same task circumstances in previous days followed by exposure to (visually) similar task circumstances that require different behavior.
2. **Insights into how big data* of actual human behavior (as captured by sensors) can be used in combination with human behavior expertise to identify and answer other questions about human behavior in the future:**
   1. Analyzing the data with a focus on identifying discrepancies in the expected amount of variation in the behavior per factor versus the actual amount of variation leads to the identification of relevant research questions on human behavior 2. The task of using big data is itself a complex task which should also be considered from a Human Factors perspective** to decrease the chance that errors occur during the use of big data. Theory on cognitive biases and automatic activation can be used to identify pitfalls in complex tasks such as these. In his PhD-research, five pitfalls were identified to be aware of.

* Big data: data for which the quantity is too large to collect and process via traditional technology.

** From a Human Factors perspective, human errors are not seen as an indication that there is something wrong with the individual who made the error but rather as an indication that parts of the system can be improved. Human Factors experts use the knowledge of human strengths and limitations, both mental and physical, to improve human-system interaction in such a way that safety, performance and/or employee satisfaction is enhanced. When trying to investigate why human errors occur, the first avenue is thus always to consider what factors within the system contributed to the causation of this error and how the system can be improved to prevent errors from occurring the future.

In the following sections in this introduction, you will read about the steps I took within my PhD research in chronological order. Firstly I used big data to identify a relevant research question on human behavior, secondly I performed the scientific research on that human behavior topic.

The chapters after this introduction are however in a different order. I will discuss the scientific research first, followed by the chapters on lessons learned about using big data of actual human behavior. I have decided to change the order of the chapters because I noticed that the original (chronological) sequence of chapters lead to too much confusion about the research question of my PhD and its scope.

Hopefully now it will be more clear that this is a dissertation on human behavior with two parts:

- The first part sharing the scientific knowledge gained on human behavior
- The second part sharing lessons learned on how big data can be used in combination with Human Factors expertise to identify and answer other questions about human behavior in the future.

1.2 Preventing organizational accidents by increasing the scientific knowledge of human behavior

Accidents at work come at a major cost including fatalities (estimated over 300,000 annual worker deaths worldwide, 5000 in the European Union), disability and economic burden [1–3]. Large improvements have occurred over the past decades, with reductions in the number of accidents and fatalities. For example in aviation worldwide, 2500 fatalities occurred during 64 accidents in 1972, whilst in 2017 there were forty fatalities during ten accidents, despite a tenfold increase in the number of flights.

1.2.1 Organizations play a role in accident causation

The ideas around accident causation have also changed over time from describing accidents as 'acts of gods' and the fault of individual employees to accidents being the result of an interaction between technical, human and organizational factors [4]. As Reason puts it eloquently in his book on organizational accidents published in 1997 [5, p.2]:

> *'Organizational accidents may be truly accidental in the way in which the various contributing factors combine to cause the bad outcome, but there is nothing accidental about the existence of these precursors, nor in the conditions that created them.'*

It is still recognized that an error of an employee at the 'sharp end' of the system can cause adverse effects, but the behavior of the employee is (to quote Reason again) "now seen more as a consequence than as a principle cause". The role and responsibility of organizations in preventing accidents is being recognized. This prevention can include eliminating error-promoting factors or adding safety barriers to prevent errors from leading to accidents.

Sometimes accidents are prevented by the proactive, off-the-cuff intervention of employees. Whilst this shows the strengths and capabilities of employees, organizations should not solely count on employees to correct their own errors or intervene proactively to prevent a system malfunction from leading to an accident. There might be scenarios in which it is not possible for an organization to introduce other measures. In these cases, the organization should make sure factors are in place that increase the probability of the intervention behavior occurring, for example by making sure there is time for recognition and response and by providing training or practice opportunities. Therefore, also from this perspective, it is important to consider the role of the organization in creating the environment that reduces accident causation and enables accident prevention.

1.2.2. We need to understand human behavior to improve systems

The previously mentioned reduction in the number of accidents in aviation illustrates that many improvements have already been made, but one can also use the same numbers to argue that not enough improvements have been made. We have not reached zero yet.

One might argue whether the goal of zero accidents is attainable, but as long as preventable accidents still occur, there is motivation to improve.

As employees interact with the system, interventions can be aimed at employees and at the system. When investigating possibilities to improve the system, it is important to take into account how the employees interact with the system and what the employee behavior is.

But what behavior should we anticipate? When it comes to the behavior and characteristics of machines and materials we can often identify a number of scenarios that can occur based on previous knowledge of the parts that something is built from and based on other machines that were built in the same way. But a human being cannot be deconstructed or recreated in the same manner.

In many ways, we are comparable to a black box with the additional difficulty that there is the perception that there is a clear causal mechanism: 'we do what we want to do'. But reality and research have proven the matter to be more complicated. Our behavior is not only influenced by our intention, but also by our environment. These context factors and their influence is not always obvious. Even in hindsight, it is not automatically clear why someone behaved the way he or she did. In order to implement successful interventions, we need to be able to predict human behavior and in order to do that, we need to understand human behavior.

1.2.3. Analyzing big data of actual employee behavior can lead to valuable insights into human behavior, íf done effectively

The social sciences have studied human behavior in a wide range using different methodologies including observation, ethnographic studies, grounded theory approaches, surveys and experiments. Also within industry, multiple methods are being used to investigate safety related issues, including incident analyses, interviews, expert judgement, ethnographic studies simulator studies. Each has their own strengths and weaknesses. The advances in technology, specifically in sensors, data collection, data storage and data analysis, make it possible to add a methodology: analyzing the actual process related behavior of employees as captured by sensors that are already present or can be embedded within the task environment.

Using (big) data of actual employee behavior as captured by sensors has multiple advantages. Whilst some of the advantages in using this type of big data are not unique to this methodology, the combination of advantages does provide unique opportunities for research. Advantages of using big data of actual employee behavior as captured by sensors includes a natural environment (in contrast to a laboratory setting) and the ability to investigate very specific hypotheses thanks to the large amount of data.

There are also phenomena that might be nearly impossible to capture with other methodologies. Asking participants or employees about their behavior is for example less reliable when it concerns automatic behavior because behavior that requires less attention tends to be encoded less well in memory. Additionally, large amounts of exposure might be required to measure an effect whilst this large amount of exposure is not feasible in an experimental or simulator setting. Finally, this type of data provides the opportunity to proactively monitor daily operation and the effect of interventions, making it possible to intervene before accidents occur.

The use of big data within organizations thus has a lot of potential benefits, but these are not automatically reaped. Big data usage on its own, as with all methods, has its challenges and using big data to investigate human behavior introduces specific

challenges. There is often a higher need for certainty with respect to the outcomes of predictive models and a need for insight into causal mechanisms. There are also often many factors that can have an influence on human behavior of which some might not be measured because there is no data on it. Other factors might require a specific combination of variables which must be included explicitly and thus require in-depth knowledge about human behavior.

Wang and Wang published an overview paper on big data in safety in 2021 where they stated: "So far, the data volume of various industries has been very large, but only a few enterprises or departments have applied big data to solve safety problems." [6] One of the limitations the authors identify is the gap between big data as obtained and the valuable safety information that is needed to obtain safety knowledge. Knowledge and skills in safety are needed to transfer the big data to valuable safety information. In order to effectively use big data within safety, big data should not simply be seen as a method applied to a topic. It is advised to integrate the discipline of big data with safety science theories to create new safety sub-disciplines [6,7].

1.2.4 Big data from Dutch rail was used in this PhD-research to obtain safety related insights on human behavior

An opportunity arose to use big data within the context of Dutch rail to study human behavior in a safety context.

The case: the influence of rail infrastructure on train driver behavior and the probability of a deceleration error

ProRail, the Dutch Infrastructure Manager, aims to reduce the amount of SPADs within Dutch rail. SPADs, or Signal Passed at Danger events, are incidents where a train passes a red aspect without authorization. SPADs receive a lot of attention within the rail industry because they can lead to severe consequences in the case of a train collision.

In 2019, there were 142 SPADs in the Netherlands [8]. Most SPADs are however not high-risk events and do not lead to any form of injury or even damage. SPADs that are non-harmful in terms of injury and damage can however still have direct and indirect costs [9,10]. But even apart from any negative consequences, SPADs are deviations from the process as intended and thus it is important for (Dutch) rail as a sector to know why these accidents occur and how they can be prevented.

New technical systems like the European Rail Traffic Management System (ERTMS) have been implemented and are expected to provide additional SPAD prevention [11]. However, in the Netherlands, the implementation of ERTMS on a national level may take up to 30 years and safe implementation can be complex, amongst others due to instability of specifications and integration issues due to different system versions and different train and trackside system suppliers [12,13]. At the same time, Dutch rail is predicted to become busier and busier. The Infrastructure Manager aims to support a growth in passenger transport of 30% by 2030 and 45% in freight transport by 2030 [14,15]. This increase requires changes to the infrastructure and the timetable.

In order to keep on improving the level of safety and improve overall performance, it is useful to understand what factors influence train driver behavior and cause SPADs. Whilst ProRail does not have any trains nor employs any train drivers, the infrastructure manager does have an influence on train driver behavior via the infrastructure design.

At the start of this PhD research, ProRail, together with the Netherlands Railways (NS), had just started using data on train driver deceleration behavior that was available as a by-product of data recorded for train maintenance purposes.

Step 1. Analyzing the discrepancies between expected and actual level of variation in behavior to identify relevant (human) factors

One of the avenues to explore was how to perform the initial analyses on the data in order to develop new insights that would provide the basis for follow-up research. Within psychology there has traditionally been a strong focus on analyzing differences averages (mean or median). In chapter 5, the Shewhart perspective is described which advocates analyzing the data from a variance perspective. This way of exploring the data in combination with human factors expertise, lead to the identification of our research question around the influence of previous exposure to signal aspects on deceleration error.

Step 2. Decreasing the probability of errors during big data utilization by being aware of possible cognitive pitfalls in complex tasks such as data verification

Human Factors specialists using big data to investigate human factors topics should be aware that they themselves are also susceptible to making errors. This raises the questions: can 'we' (Human Factors experts) also use our knowledge of human factors to decrease the probability of errors during the use of big data?

In our research, we encountered one of the common challenges within big data, namely suboptimal data quality. Improving the data quality was important because we wanted to use the data to understand the contributing factors to human behavior and use this information to potentially implement costly interventions, thus requiring a high degree of certainty. Data verification is a task that can be affected by human error. In order to improve the data verification process, it is important to take cognitive biases into account. Chapter 6 describes the cognitive biases that can occur among researchers and analysts and negatively influence the verification process.

> It should be noted that there are standards within railways that ensure that safety critical software is made as safe as possible. The data that was used in this research was not part of any software with a safety critical component. There are also many different ways in which data verification can be performed. Chapter 6 gives insight into pitfalls that can be present during the data verification process of data that is available but does not yet comply to the strict standards applied that are relevant for safety critical software.

Other big data related tasks such as result visualization and interpretation can also be examined for the presence of common pitfalls. This is beyond the scope of this dissertation, but the principles applied in chapter 6 can be used to also inspect other tasks within big data utilization from a Human Factors perspective.

Step 3. Deciding which (human) factor to investigate further

Once the data was improved, analyzing the data from the Shewhart perspective (as described in chapter 5) pointed towards a potential cause of human error that might otherwise not have been given a lot or any attention, namely the role of incidental learning or the non-intentional learning that automatically occurs during daily interaction with the infrastructure. In section 1.2.5. of this introduction, incidental learning is further introduced.

The potential impact of incidental learning was chosen as an ideal candidate to study as part of the main research question of this dissertation because:

- which has not been investigated yet in the context of human error
- which can be influenced by an organization in line with the Human Factors perspective of improving systems to reduce the probability of errors and increase safety, performance and employee satisfaction
- which is difficult to study using other methodologies
- which relates to fundamental human psychology and thus increased knowledge about this factor is not only beneficial for the railways but for all industries

Step 4. Investigating the impact of the factor on the behavior and remaining safety margins

The impact of incidental learning was first analyzed using behavioral data as the dependent variable. This rich data source allowed the testing of nuanced hypotheses. The specific analysis used in this study is described in detail in chapter 2.

Big data can be analyzed using a 'bottom-up' approach where for example machine learning is used to find the most significant variables. We combined big data with a 'top-down' approach guided by content expertise, in this case human factors knowledge and theory about incidental learning, to:

- Improve the measure of human behavior
- support the data verification by identifying unlikely behavior patterns that could be caused by data quality problems
- determine which factors to include in analysis
- determine how the factors should be operationalized
- decide which additional factors to include in subsequent analyses (see step 5)

The calculation of multiple factors in the studies described in chapter 2 and 3 required such a specific combination of variables that it is unlikely to have been found by a big data analyst or team of analysist without thorough interaction with experts on human error.

Step 5. Investigating the impact of the factor on incident occurrence

The knowledge gained in step 4 was used to decide how to design the follow-up research where incident data was used as the dependent variable. This data source makes it possible to study whether an error also actually leads to accidents. Multiple factors can influence whether an error will indeed lead to an incident or could be corrected in time. To expand the knowledge of incidental learning of error and incident cause, we considered the role of self-correction. We calculated the 'opportunity for correction': the amount of room present for self-correction by the train driver as influenced by the infrastructure. The specific analysis used in this study is described in detail in chapter 3.

1.2.5 Incidental learning as a cause of train driver error

Incidental learning is by definition not positive or negative. Incidental learning is the learning that occurs without explicit intention [16]. It is the on-the-job learning that

occurs, in contrast to learning during training sessions and courses. In experimental settings, intentional learning and incidental learning are differentiated by the instruction that participants are given. During the incidental learning condition, the participants are not aware of the learning situations and are not instructed as to what they will truly be tested on [17].

Within research on incidental learning, multiple studies can be found on the topic of language acquisition (such as [18–21]). Other studies on incidental learning that also focused on information acquisition use simple cognitive tasks such as word recall and recognition (such as [22,23]) or self-reported levels of incidental information acquisition ([24]). Models on incidental learning have also been proposed that take a broader, conceptual view of incidental learning in technology use ([25]) and learning within organizations (such as [26]). These studies do not describe the effect of incidental learning on specific behaviors and when focused on behavior in general, they center around the positive effects of incidental learning.

In this dissertation, incidental learning is examined through the lens of influencing behavior and influencing it negatively. A specific form of incidental learning is considered, namely the strengthened neural activation for certain behavior upon perceiving a cue or stimulus. This can lead to a behavior being performed after perceiving a cue or stimulus whilst that behavior is in fact erroneous. We can call this 'the usual response bias'. In the context of a pedestrian sign, the usual response is to start walking when the sign turns from red to green whilst the usual response is to stop walking when the sign turns from green to red (before crossing). Due to the neural adaptation, there is an inclination to perform the usual behavior upon perceiving a cue, which can be suitable or unsuitable behavior in a given situation. When the behavior is unsuitable, an error occurs. This can then be called the 'usual response bias' as a negative result of incidental learning. When the behavior is suitable, this is the positive result of incidental learning.

If there is indeed a significant negative influence of incidental learning, this is important to understand as it can undermine results of explicit training and awareness campaigns (See **Figure 4**). It is of course important to train employees (top right of **Figure 4**), but if incidental learning teaches employees different behavior (bottom left of **Figure 4**), then this explicit training will be partly undone.

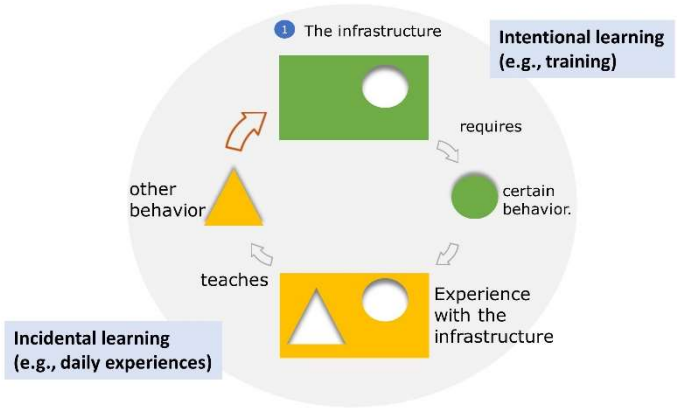


Figure 4. Differences in learning during intentional learning and incidental learning can lead to human error despite thorough explicit training. The green resembles an employee receiving explicit training on how he or she should perform according to company standards. The yellow resembles what an employee actually experiences on a day-to-day basis. If these differ, then human error can occur despite the employee successfully following the intentional learning sessions.

The two main research questions with respect to incidental learning are:

- Can incidental learning negatively influence human behavior?
- If the answer is yes: under which task design circumstances does this occur?

In chapter 2, incidental learning is described in depth and the train driver behavior is used to test whether incidental learning can indeed negatively influence behavior and under which frequency of exposure and (visual) similarities. In chapter 3, the results are validated by examining incidental learning again, but this time by using incident data. An additional factor is also introduced in chapter 3, namely: what about the train driver's ability to correct their initial error?

The research described in chapters 2 and 3 adds to the current body of science on incidental learning which is mainly focused on language acquisition and uses contrived settings in the research design. Based on the results, we also recommend an expansion of the field of Human Factors to not only consider what the work-environment of an employee entails at the moment of performing a task but also what they were exposed to in the past. Additionally, the data-based approach gives us a firmer grasp on the extent of the negative influence of incidental learning. When it comes to human behavior, many behaviors can have an influence. The more relevant question is therefore often not: can a factor influence human behavior and error occurrence, but to what degree does a factor increase the probability of human error and under which circumstances?

Specifically for the rail industry, this research on incidental learning provides concrete do’s and don’ts around infrastructure design to reduce the occurrence of SPADs. A practical strength of the data-based approach lies in both identifying the infrastructure and scheduling designs which increase the probability of the driver committing an error, ánd in identifying designs which do not increase the probability of the driver committing an error. Information on which factors or circumstances do not require an intervention is also important. The data-based approach allows us to zoom in and see where the problem lies exactly so interventions can be made more specific and in congruence with other organizational goals such as productivity and quality.

1.3 Reading notes to each type of reader

To the reader who has affinity with Human Factors and data: We are two peas in a pot. Whether you work in academia or in industry or a combination of both: I hope this work will inspire and enthuse you the same way it did me.

To the reader who is head of the safety department in a (high-risk) industry: Reading this work will show you what is possible when data and psychology expertise are combined. If you want to reap similar benefits within your own organization, I would advise:

A. recommending this work to someone within your team that has affinity with Human Factors ánd data, or
B. recommending this work to someone who understands the merits of combining expertise and knows how to build a team that brings these experts together in a setting where mutual learning is promoted and adopted. Proactive effort is needed to build a shared bridge between the 'data minded' and the 'psychology minded'.

To the reader who is interested specifically in learning more about incidental learning as a cause of human error and how to apply this knowledge to prevent accidents: I recommend chapters 2 through 4, and especially section 4.2: 'Concrete application of the insights on incidental learning'.

To the reader who is interested in SPAD prevention: Chapters 2 and 3 will give you the information and evidence for one cause of SPADs, namely incidental learning. In chapter 4, questions are listed that can be used during incident investigation to consider whether incidental learning might have played a role. If you are interested in performing more research on SPADs using data on train driver behavior, then the information on our proactive SPAD indicator can also be useful. This information can mainly be found in chapter 2, with tips to prevent accidental erroneous use in chapter 6, tips on how to analyze the data to identify other relevant behavior influencing factors in chapter 5 and additional tips on how this measure can be used within an organization in chapter 7.

Please note that this is a dissertation about human behavior irrespective of the industry. It is thus not meant as an extensive exploration to the causes of SPADs. Many factors can be involved in the causation of SPADs including problems with the train, tracks or signal aspect functioning, communication between train driver and dispatcher, suboptimal working conditions (within the cabin) and suboptimal signaling design. This dissertation only focuses on the possible influence of suboptimal signaling design. Within signaling design, there can also be multiple problems including the topic of this dissertation but also factors play like visibility and how clearly it is communicated which signal belongs to which track.

For those who want to read more about other SPAD causes or about other ways to use big data within railway safety, I have listed some suggested reading in appendix E although this list is by no means exhaustive.

As a final notes about Dutch rail as a sector: Since I studied human error among train drivers and the impact of the infrastructure on 'eliciting' these errors, my texts focus on where there is room for improvement within Dutch rail. My work is however by no means a value judgment on Dutch rail or any of the employees working (directly and indirectly) to keep the trains running safely and on time. If I were to cast a verdict from a personal perspective, it would be a positive one. During all the years working with my rail colleagues I have only encountered people that also wanted to contribute to improving the railways and made effort to do so.

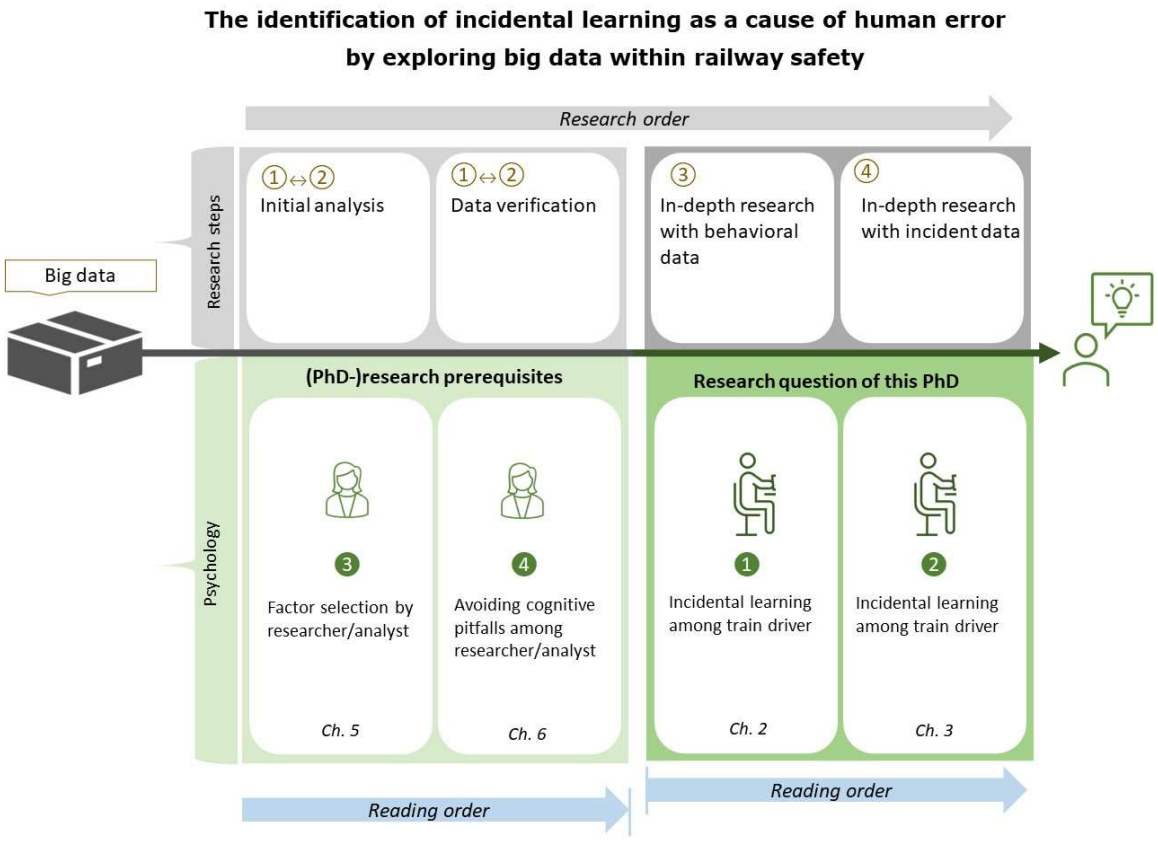


Figure 5. Overview of main dissertation chapters showing the difference between chapter/reading order (blue arrows at bottom) and the order of the research steps when performing Human Factors big data research (grey arrow at the top). Chapters 2 and 3 will answer the research question of this PhD. Chapter 4 contains the conclusions and recommendations around incidental learning based on chapters 2 and 3.

Chapters 5 and 6 describe part of the prerequisite steps in order to use big data to gain insights such as those obtained around incidental learning. Chapter 7 contains additional recommendations around using big data to improve safety within organizations.

Chapter 2.

Train driver behavior is influenced by incidental learning

Based on the article: Burggraaf J, Groeneweg J, Sillem S, van Gelder P. What Employees Do Today Because of Their Experience Yesterday: How Incidental Learning Influences Train Driver Behavior and Safety Margins (A Big Data Analysis). Safety. 2021; 7(1):2. https://doi.org/10.3390/safety7010002

Chapter summary

Employee behavior plays an important role in the occurrence and prevention of incidents by affecting safety margins. Within Dutch rail, train driver deceleration behavior influences the safety margins with respect to Signal Passed at Danger events (SPADs). In this chapter and chapter 3, we examine the potential impact of incidental learning on train driver behavior. Incidental learning is the day-to-day on-the-job learning that occurs unintentionally. This learning influences which behavior (schema) is more likely to be activated in the employee's brain. We focus specifically on the incidental learning that occurs in the presence of variation in task design.

We posit that:

- if employees are frequently exposed to a task that requires a specific behavior,
- ánd there is a different task requiring different behavior but with a (visually) similar cue,

then the probability of an error due to activation of an unsuitable behavior increases (See **Figure 6)**.

The Dutch rail system has variation in task design, namely in the yellow aspects shown at the same signal and in the signal distances. This leads to different behavior being required at different times with differing levels of (visual) similarity between the task that require different behavior.

In the study presented in this chapter, we used behavioral data as captured by sensors on passenger trains in the Netherlands. The train driver deceleration behavior during a red aspect approach was summarized in an indicator called mDtSPAD. The analysis included 19 months of data that was already being captured. Human Factors expertise was used to determine which variables needed to be calculated using the raw data in order to effectively investigate the impact of incidental learning. For the statistical analysis we used a variation on piecewise regression to gain more insight into the exact shape of the relation between frequency of exposure to signal aspects in the previous 14 days and error probability.

The analysis showed changes in behavior when the train drivers had been previously exposed to different behavior requirements in the same location with a similar yellow aspect. These results imply that task design can be improved by taking into consideration what an employee is exposed to during other moments of the shift, and not just during the execution of the specific task.

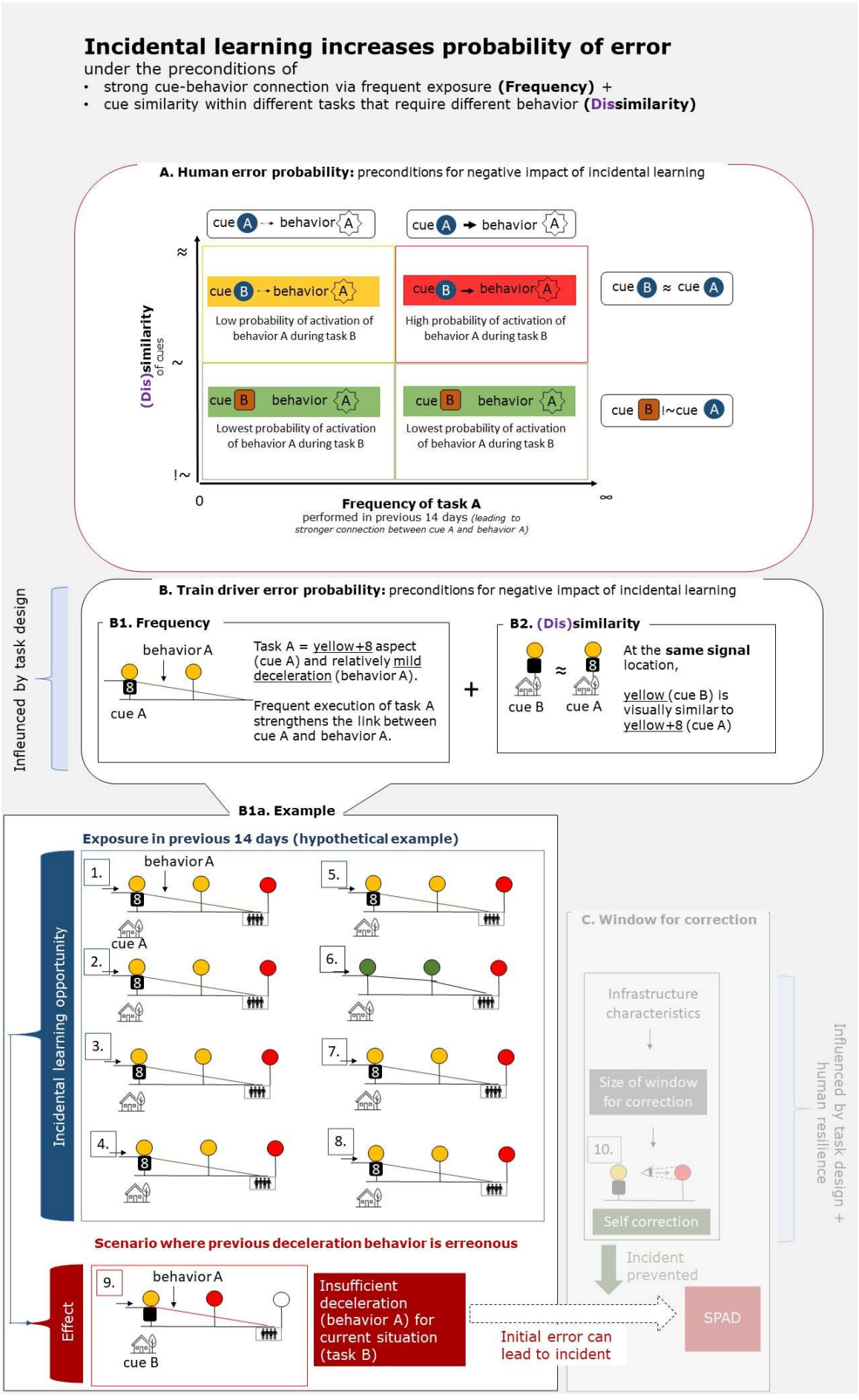

Incidental learning increases probability of error
under the preconditions of
• strong cue-behavior connection via frequent exposure (Frequency) +
• cue similarity within different tasks that require different behavior (Dissimilarity)
A. Human error probability: preconditions for negative impact of incidental learning
cue A behavior A
cue A behavior A
cue B behavior A
Low probability of activation of behavior A during task B
cue B behavior A
High probability of activation of behavior A during task B
cue B ≈ cue A
cue B behavior A
Lowest probability of activation of behavior A during task B
cue B behavior A
Lowest probability of activation of behavior A during task B
cue B !~cue A
(Dis)similarity of cues
≈
~
!~
0
∞
Frequency of task A
performed in previous 14 days (leading to stronger connection between cue A and behavior A)
Influenced by task design
B. Train driver error probability: preconditions for negative impact of incidental learning
B1. Frequency
behavior A
cue A
Task A = yellow+8 aspect (cue A) and relatively mild deceleration (behavior A).
Frequent execution of task A strengthens the link between cue A and behavior A.
+
B2. (Dis)similarity
cue B
≈
cue A
At the same signal location,
yellow (cue B) is visually similar to yellow+8 (cue A)
B1a. Example
Exposure in previous 14 days (hypothetical example)
Incidental learning opportunity
behavior A
cue A
1.
2.
3.
4.
5.
6.
7.
8.
Scenario where previous deceleration behavior is erreonous
Effect
9.
behavior A
cue B
Insufficient deceleration (behavior A) for current situation (task B)
Initial error can lead to incident
C. Window for correction
Infrastructure characteristics
Size of window for correction
10.
Self correction
Incident prevented
SPAD
Influenced by task design + human resilience

Figure 6. Summary of incidental learning with the most important rail related hypothesis in chapter 2 used as an example. This figure is the same as **Figure 3** at the end of the summary, with only section C blurred. Section C is blurred here because this element is investigated in chapter 3.

Section A illustrates under which preconditions incidental learning can lead to an error. Consider that correct execution of 'task A' consists of perception of 'cue A', followed by performance of 'behavior A', and correct execution of 'task B', consists of perception of 'cue B' followed by performance of 'behavior B'. Incidental learning can lead to an error during task B when 'behavior A' is activated instead of 'behavior B' after perception of 'cue B'. This can occur IF

- 'cue A' and 'cue B' are similar ánd
- perception of 'cue A' strongly activates 'behavior A' as a result of frequent exposure to and execution of 'task A'

Section B shows one example of how these preconditions can take place in Dutch rail, using the example of signal aspect 'yellow+8'.

The signal aspect yellow+8 is 'cue A', with a mild deceleration as 'behavior A'. The signal aspect yellow is cue B with 'behavior B' being a more stronger deceleration in order to stop in front of the red aspect.

Section B2 visualizes the similarity between signal aspect yellow+8 and yellow at the same signal location. **Section B1** visualizes the frequency of task A which influences the strength with which perception of yellow+8 (cue A) activates the mild deceleration (behavior A).

Section B1a gives a simplified example of a train driver being exposed to and executing task A seven times (approaches 1-5 and 7-8). During approach 9, the yellow aspect (cue B) is present as part of task B with the correct behavior being a stronger deceleration (behavior B), but mild deceleration (behavior A) is performed which is an error in this setting.

2.1. Introduction

When SPADs occur, sometimes the cause is easy to identify, but on other occasions it remains unclear what the exact cause was. This is especially the case when SPADs do not occur because of a technical failure but because a mistake was made in train driver behavior. Whilst "insufficient deceleration" is often identified as a cause of a SPAD, it is harder to identify why the train driver did not decelerate sufficiently [27]. In other words: what factors created the situation that caused the train driver to not decelerate sufficiently? Different experts have also been shown to regard different factors as having the greatest influence on the occurrence of the same incident, based on their interpretation of the same incident report [28,29].

The field of Human Factors looks into the factors that influence human behavior and can increase or decrease the probability of an error [30]. Human factors research in rail has become more common since the mid to late 1990s [31].

One factor that is often considered within rail is visibility of signals. Another example of a factor is signal placement, with a focus on whether it might be confusing for the train driver to know which signal is for him/her [31–35]. Route knowledge can help to prevent mistakes as a result of problems in signal visibility or interpretation [36,37]. Another option is to improve placement and make the infrastructure more logical from the train driver's perspective.

Improving the placement of signals is a more fundamental solution in contrast to route knowledge, where there is more variability with respect to its implementation (e.g. when the driver last drove in that location and the level of knowledge that was gained and will be sustained under stressful situations)[38]. Infrastructure changes are a more reliable safety barrier than interventions related to training and awareness. It therefore makes sense to consider whether the infrastructure can be adjusted to decrease the probability of human error.

In rail, the infrastructure is an important part of the working environment of the train driver. We thus advocate improving that working environment. Often, the effect of the working environment of the train driver at the moment of his/her task execution is considered. This is a core question within human factors: "which factors (at moment x) influence the performance at moment x?" And often in incident analysis: "which factors (at moment x) caused the error at moment x?" But what about the past? Does it matter what an employee experienced yesterday or last week or last month? In other words: "do factors (at moment x-Δt) cause an error at moment x?"

In this chapter, we examine incidental learning as a factor impacting human behavior [39]. Incidental learning can have a positive or a negative impact. Incidental learning is the learning that occurs without an explicit intention [16]. In experimental settings, a distinction between intentional learning and incidental learning is made depending on the instructions that participants are given. During the incidental learning condition, the participants are not aware of the learning situations and are not instructed as to what they will truly be tested on [17].

Incidental learning is part of the on-the-job learning that occurs, in contrast to learning during training sessions and courses. People learn every day, both intentionally and incidentally. For skilled work, a lot of learning takes place on the job. A common saying is "you learn by doing". This learning does not stop once we are able to perform a skill. Learning is the creation of new pathways in our brain and also the strengthening of existing pathways [40].

Incidental learning is difficult to identify as a cause for changes in human behavior. One reason for this is that incidental learning can be part of implicit learning. This means that the employee is not necessarily aware of what he or she has learned or even that he or she has learned. Implicitly learned knowledge can control action, but the learner himself is not able to tell others that this is what happened [41–43]. Wang and Theeuwes focus on implicit attentional bias and show that people quickly pick up on visual changes in the environment and change their behavior accordingly even though they are not aware of the changes. They conclude that "people adapt to a changing environment but that there are lingering biases from previous learned experiences that impact the current selection priorities" [44].

Another reason that incidental learning can be difficult to identify is that during an incident analysis, the situation at the time of the incident is analyzed. Whilst the causes of the situation might also be analyzed, the preceding "normal" situation is often not analyzed. Thus, what the employee or train driver is exposed to on a daily basis before the incident is not necessarily considered. Even when it is, it is hard to prove the impact of previous exposure, i.e., incidental learning. In the case of SPADs, there are simply not enough incidents to analyze this cause systematically without specific direction and detailed hypotheses. A third reason for difficulty of detecting incidental learning in the past may be small effect size.

2.1.1. Incidental Learning Influences the Schemas in an Employee's Brain

Incidental learning influences the development and activation of schemas in an employee's brain. Schemas embody the procedural knowledge that is needed to carry out actions [45–47]. Schemas can be described as generalized procedures for carrying out actions. In novel tasks, when a schema does not yet exist, much attention is needed to carry out the action. Once schemas are present, these actions can mostly be performed automatically, i.e., with little attention required. Schemas thus help us perform actions more efficiently [48]. Actions will be performed correctly if the right schema is activated at the right time.

Schemas can be activated in a top-down fashion via the intention to perform an action. This requires attention. Schemas can however also include triggering conditions. If the environmental conditions match the triggering conditions, then the schema can be activated without conscious thought. For example, if one has a cup nearby on the desk, he/she can pick it up and have a sip without explicit intention or even thirst. The mere sight of the glass can trigger the schema to pick it up (see Ref. [49,50] with respect to unconscious control of motor action; Ref. [51–54] specifically for hand movement).

An event (a cue) can become a trigger for a schema when it is often paired with the execution of the schema. The more often they are paired, the stronger the schema activation will be upon perception of the cue. This linking of a cue to a schema is part of incidental learning.

Problems occur when the incorrect schema in one's head is activated. Correct behavior is then activated, but it is unsuitable for the specific situation. We hypothesize that this is more likely to occur if there is variation in task design. Specifically, we posit that human error is more likely to occur if different behavior is required in (visually) similar settings.

An example is crossing the street on foot. In right-driving countries, pedestrians should look left and right and left again, before crossing. When a pedestrian goes on holiday to a left-driving country, he or she should look right and left and then right again, but the pedestrian is inclined to look in the pattern he or she is used to, namely left–right–left.

This is clearly not caused by a sudden lack of head turning ability, but caused by a different requirement in a similar situation (crossing a road). It can therefore occur even if the pedestrian is fully aware of the rules that apply in a given country and wishes to adhere to them (see e.g., research using the Stroop test for ample evidence of people erring in the simple task of naming a color because they read the colored word instead [55]).

The same applies to driving a car. People are perfectly capable of taking a roundabout clockwise. They are also perfectly capable of taking a roundabout anti-clockwise. However, going on holiday and driving on the opposite side of the road than one is used to is very difficult the first few times. When there are other cars around, this is a visual reminder that one is in a different country and the roundabout should be taken the other way round. However, when there are no other cars in sight or there are other distracting traffic situations present, it is easy to veer into the old pattern and take a roundabout the wrong way round.

2.1.2. Incidental learning in Rail

During a red aspect approach, it is the train diver's task to decelerate sufficiently to stop in front of the red aspect. The driver has schemas in his brain for the deceleration behavior. These schema's can be activated by the signal aspects along the tracks or other cues.

The signal aspects along the tracks provide information on which behavior is suitable. A red aspect in Dutch rail is always preceded by a yellow aspect to inform a train driver that a red aspect is coming and that he should start to decelerate. In contrast to road transport, this is necessary because trains have a very long braking distance (e.g., 580 m at 140 km/h and an emergency deceleration of 1.3 m/s^2, not taking driver reaction time, reaction time of the brakes and train-track friction conditions into account).

In the Dutch signaling system, the aspect sequence green–yellow–red is most common, but other yellow aspects are also used. For example, if the distance between the yellow and red aspect is not long enough given the speed of the train, then the yellow aspect will be preceded by another yellow aspect. This yellow aspect can be plain yellow again, but it is most often yellow in combination with a number.

Yellow+8 for example indicates that the train driver needs to drive 80 km/h at the next signal, and if that signal shows a yellow aspect, the train driver knows that the following signal is red and (s)he needs to decelerate to be able to come to a stop in front of the next signal (see **Figure 7**).

Figure 7. Red aspect approach preceded by yellow aspects

There are multiple forms of variation in rail task design that can cause incorrect schema activation after incidental learning. One type of variation is the combination of variation in permitted track speed and in distance between signals. These cause variation in the amount of deceleration that is necessary to stop in front of the red aspect. In **Figure 8** it

is illustrated that in the left scenario, a continuous deceleration rate of 0.26 m/s^2 would be sufficient to stop in front of the red aspect, while in the situation on the right, a deceleration rate of 0.59 m/s^2 is needed. If a driver is more often exposed to the situation on the left, then the cue "yellow aspect" can trigger the initiation of a schema resulting in a slower rate of deceleration than required for the situation on the right.

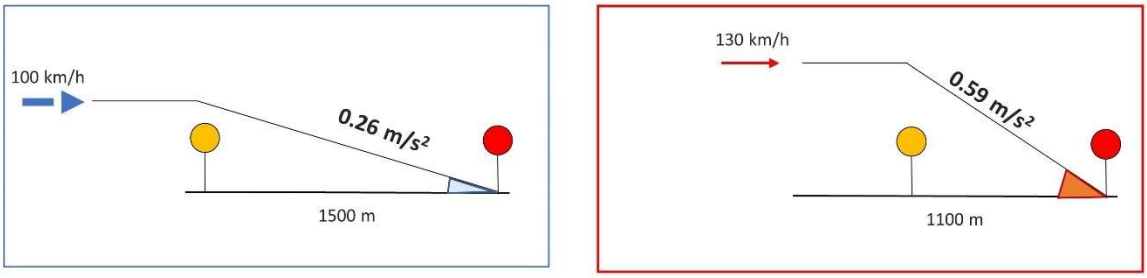


Figure 8. here is variation in the necessary rate of deceleration. In the left approach, a continuous deceleration rate of 0.26 m/s2 is sufficient to stop in front of the red aspect, while the approach on the right requires a deceleration rate of at least 0.59 m/s^2.

The above example illustrates variation in the required deceleration for the same signal aspect (yellow). In Dutch rail, there is also variation in which signal aspect is present at a given location. As described in section 2.3, the yellow aspect can be preceded by other yellow aspects such as yellow with the number four (yellow+4). **Figure 9** shows a location where signal Sx can have signal aspect yellow+4, as part of a yellow+4-yellow–red sequence. It can also have a yellow aspect as part of a green–yellow–red sequence as displayed in the bottom scenario with the previous green aspect not shown in the figure.

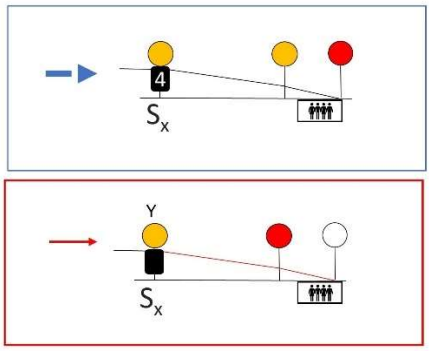


Figure 9. There can be variation in the signal aspect at a specific location. In the top approach, the first signal has aspect yellow+4 because the signal at the station is red and the distance between the last two signals is insufficient for a green–yellow–red sequence. In the bottom approach, the first signal is yellow because the next signal is red.

A signal can also have a yellow+number aspect as part of a speed restriction. This kind of speed restriction is sometimes needed to prevent trains from driving too fast over a switch (**Figure 10**). Aspect yellow+4 indicates a speed restriction to 40 km/h by the next signal, whilst yellow+8 signals a speed restriction to 80 km/h, etc.

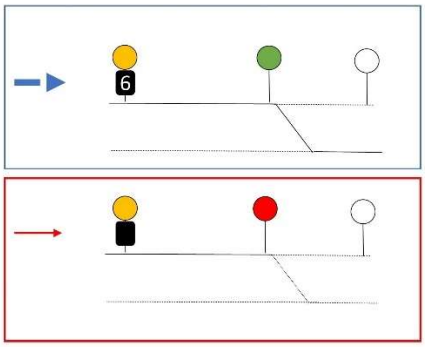


Figure 10. There can be variation in the signal aspect at a specific location. In the top approach, the first signal has aspect yellow+6 because the switch after the next signal has a maximum permitted speed of 60 km/h. In the bottom approach, the first signal is yellow because the next signal is red.

In this study, we investigated the effect of the above variations on train driver deceleration behavior. An additional infrastructure characteristic that was taken into account was the track speed limit just before the first yellow aspect. We did not expect to see an effect of incorrect schema activation for those approaches with such low speed that the train driver could start to decelerate upon sight of the red aspect and still come to a standstill with mild deceleration. The issue of incorrect schema activation is assumed to be mostly relevant for those approaches where the train driver needs to decelerate before the red aspect is visible. This is because automatic behavior can also occur without the use of schemas. This can occur when the information needed to perform the action is directly available in the environment [56]. When the driver can see the red aspect, he can estimate the distance and the best rate of deceleration. When the train driver has to start decelerating before seeing the red aspect, he needs to rely fully on the information stored in the schema in his long-term memory.

2.1.3. Previous Research on Incidental Learning and Task Design Variation

The field of human factors looks at the influence of system or task design on human behavior [30]. There is however a strong focus on task design at the moment of performing the task and not on the potential influence of previous exposure to other task designs. Experience is also mentioned as a positive factor, without the nuance that experience in combination with task design variation can lead to errors.

Some commonly used taxonomies of human error causes, for example, do not include this factor. One accident analysis method called the Human Factors Analysis and Classification System (HFACS) was inspired by Reasons' popular Swiss Cheese model and provides a taxonomy of failure across four organizational levels: unsafe acts, preconditions for unsafe acts, unsafe supervision, and organizational influences [57]. Of the seven preconditions for unsafe acts, the "technological environment" is most aligned with the idea of task design. This precondition is further clarified as encompassing "a variety of issues including the design of equipment and controls, display/interface characteristics, checklist layouts, task factors and automation" (p.62). The focus is mostly on the state of the individual at that precise moment, and not the impact of previous learning.

One human reliability analysis method, the SPAR-H method, estimates error probability and contains a list of performance-shaping factors (PSFs). The eight PSFs are: available time, stress/stressor, complexity, experience/training, procedures, ergonomics/human-machine interface, fitness for duty, and work process. The PSF "experience/training" can only be scored as poor, nominal, or good (or it can be considered that there is insufficient information). For this factor, more experience is considered better and reduces the (calculated) probability of an error [58]. In our research, the hypothesis is that greater experience can lead to errors, if combined with problems in task design. The PSF "ergonomics/human-machine interface" comes closest to the idea of task design, but focuses mostly on the state at that moment and not the impact of previous learning.

In scientific SPAD literature, the role of infrastructure elements is mainly considered with respect to visibility and interpretability of the signal [33–35]. One study on driver performance modeling and its practical application included line speed as related to signal and sign visibility and reading times [32]. One human factors SPAD hazard checklist contains the following scoring factors: the presence of driver's personal factors, driver inattentiveness, signal visibility, the association between the signal and the correct line, the ability to read signal aspect correctly, the ability to interpret signal aspect correctly, and the ability to perform correct action [59]. There is no factor for task variation.

Within the rail industry, there are some recommendations on infrastructure variation. The Independent Transport Safety and Reliability Regulator in Australia for example recommends making sure that there are no standard caution or low speed aspect leading up to the red aspect, because "permanent caution signals, for example, do not provide drivers with information about the next signal, and can therefore be a SPAD trap" [60]. Incident investigations at the Dutch Rail infrastructure manager ProRail have also led to the hypothesis that certain types of variation in aspects at the same signal location can pose a risk. Research including the detailed psychological mechanisms and specifically the effect sizes has however been missing. Up until recently, there was not enough data to test these effects rigorously.

The UK Rail Safety and Standards Board (RSSB) conducted a large-scale investigation in 2016, reviewing 257 industry SPAD investigation reports and organizing SPAD workshops with 60 participants with various job titles from freight operating companies, passenger operating companies, and the UK Infrastructure manager Network Rail [61]. They identified 10 risk management areas, such as signal design/layouts and driver competence management including route knowledge. The recommendations for signal design/layouts focus mostly on visibility of the signal and design of the signal itself and of the gantry. The route knowledge was considered as positive in that report. Route knowledge is also in other countries mentioned as a positive and important factor [38,62]. Variations in signal aspect shown on the same route are not mentioned.

Balfe, on the other hand, mentions expectation bias as a factor influencing SPADs in her review of 83 internal investigation reports of SPADs occurring between 2005 and 2015 on the Irish rail network [63]. The exact link between expectation bias and infrastructure is not specified. This author does mention the potential for congested networks to result in single or double yellow aspects being routinely experienced by drivers across a route, thereby leading to an expectation of continued movement rather than a subsequent stop signal upon seeing a yellow aspect.

It should be noted that the term 'expectation' is not clearly defined in the Dutch railways despite its use in accident analysis reports. Van den Top (2010) noted that in SPAD reports it was often written that the accident was caused by the driver 'having an expectancy'. He however argues that it would be more accurate to state that the driver

had 'a false expectancy' and that the focus of causal analysis should be on the system related course of events that eventually led to both the false expectancy and the lack of timely correction of that false expectancy. For a more elaborate and nuanced description of the current and desired use of words such as 'expectancy' and 'attention' and 'distraction' in Dutch SPAD analysis reports, see [38]

2.1.4. Objective of this Study

The objective of this study was to investigate whether incidental learning impacted employee task performance in the presence of task design variation. We hypothesized that incorrect schema activation caused lower deceleration rates and thereby smaller safety margins between trains and red aspects. We focused on similarities in the yellow aspect and the location as triggers for schema activation. The specific question was:

- Does frequent exposure to certain signal aspects (at certain locations) impact the behavior in a (visually) similar but deviating situation?

In previous railway research, research questions like this could not be answered due to small sample size. Thanks to technological developments, we now have different tools that make it possible to answer questions that could not be answered in the past.

2.1.5. Hypotheses

We hypothesized that incorrect schema activation can cause insufficient deceleration, potentially resulting in SPADs or near-misses. We identified four situations where this could occur. The more common signal approach is here referred to as "the standard approach". The less common approach is referred to as "the deviating approach". This deviating approach is also the safety-critical approach. If the schema of the standard approach is activated during the deviating approach, then an incorrect schema is activated. The more often the train driver is exposed to the standard situation, the higher the chances of incorrect schema activation during the deviating approach.

In Dutch rail, there are two main types of situations where a specific signal is often yellow and can become "the standard approach":

- When the scheduling is such that the signal at the train's stopping location often has a red aspect. The signal(s) preceding it will have yellow aspects at an equal frequency. We call this "yellow entrance to the station". A distinction will be made below between the 'yellow+number entrance effect' and the 'yellow-yellow-red (entrance) effect'.
- When that signal often functions as a speed limit indicator in front of a switch.

The 'standard situation' is thus the situation where the aspect is commonly yellow because of the scheduling or as a speed limit. The 'standard situation' and accompanying aspect sequence is visualized in the below figures in the blue box located at the top of the figure. The 'deviating situation' is then when one signal earlier shows a red aspect. In the same figures, the 'deviating situation' is shown in the red box at the bottom of the figure. These figures are used to show that, at the same location, there are (visual) similarities between the standard and deviating situation but the required behavior is different, because an earlier signal has the red aspect.

A location where a signal often has a speed restriction is visualized in **Figure 11**[2]. The standard, more common situation is shown at the top (blue). Below this is the deviating situation (red). In this scenario, the location is exactly the same during both approaches and the signal aspect is visually similar.

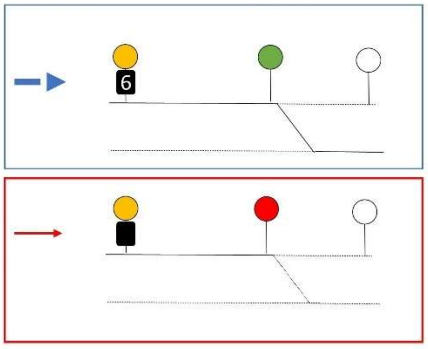


Figure 11. The yellow aspect in the deviating approach (bottom) is at the same location as the yellow+number aspect during the standard approach because of the upcoming switch (top). For both approaches the cue is visually similar (yellow vs. yellow+number) and the location is the same.

In the above example, the yellow and yellow+number signal aspects are said to be visually similar. Visual similarity is defined by the number of shared points or common features, and the type of difference. Visual similarity is higher with deletion at end points (such as the number 4 not showing below the aspect) than for differences like deletions leading to breaks in continuity or mirror image reversals [64]. The signals with yellow and yellow+number aspects are thus visually similar because they have many visually identical points with the difference being a deletion at the bottom (**Figure 12**).

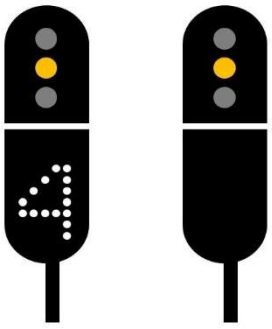


Figure 12. Signals with "yellow+number" and "yellow" aspects are visually similar.

Speed restriction and entrance at yellow can also occur at the same location (**Figure 13**). While the signal aspects can differ (e.g., yellow+8 and yellow+4), the only relevant situations are those where both are the same.

[2] The figures are simplified visualizations for explanation purposes. In reality other signals can also have a light box (visualized in the figures by the black rectangle below) to show numbers. In this study the focus lies on the aspect of the first signal in the figures.

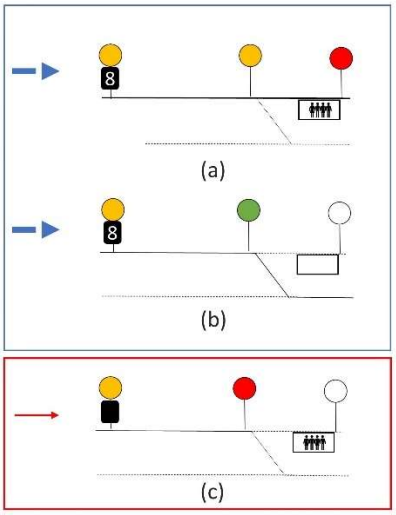


Figure 13. The yellow aspect in the deviating approach (c) is at the same location as the yellow+number aspect during the standard approach because of the entrance at station and/or the upcoming switch (a and b). During the standard and deviating approaches, the cue is visually similar (yellow vs. yellow+number) and the location is the same. Note: Both yellow+number aspects must be the same during the blue approaches.

It is also possible for entrance at yellow to occur with a yellow-yellow-red sequence (See **Figure 14**[3]). In this scenario, the signal aspect and location are exactly the same during both the standard and deviating approach. These scenarios are interesting from a theoretical perspective because the aspect at the first signal is exactly the same in both scenarios. These situations are however not common anymore in the Netherlands and especially not at locations with high track speeds (for more information on track speed filters in the analysis, see method section 2.2.2).

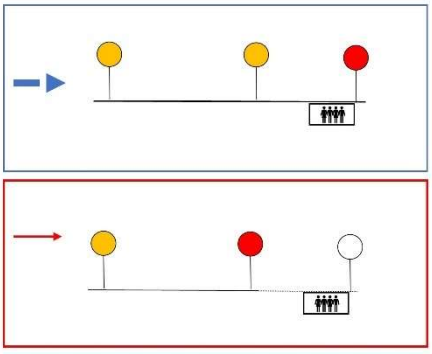


Figure 14. The yellow aspect in the deviating approach (bottom) is at the same location as the yellow aspect during standard approach towards the station (top). During both approaches, the cue is the same (yellow) and the location is the same. The station stop is shown by a rectangle with passengers.

The last hypothesis also encompasses the same aspect in both the standard and deviating situation, but with differences in location. As mentioned previously, the distances between signals varies during approaches where the red aspect is preceded by a yellow and green aspect (GR–Y–R approaches). The track speed and signal distance

[3] The rectangle with passengers is used in the figures to indicate the scheduled stopping location for the train. In these yellow-yellow-red scenarios the actual platform is however usually longer and already starts in front of the second signal with passengers potentially waiting in that location for a different train scheduled to stop there.

determine the amount of deceleration that is needed. We call this "theoretical mean deceleration". The theoretical mean deceleration does not have the same value across the Netherlands. The hypothesis is that GR–Y–R approaches with higher theoretical mean deceleration values are deviating situations in comparison to GR–Y–R approaches with lower theoretical mean deceleration values (See **Figure 15**). If the schema of the blue (left) situation is activated during the red (right) situation, insufficient deceleration is used.

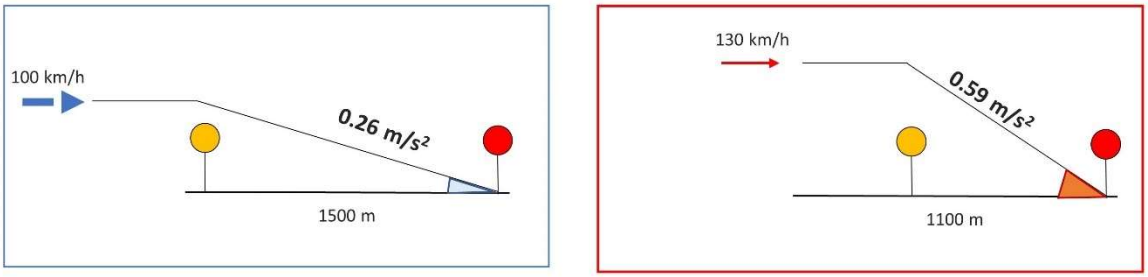


Figure 15. During the deviating approach (right), a greater deceleration rate is required than during the standard approach (left). During both approaches, the cue is the same (yellow aspect) but the location is different. The two theoretical mean deceleration values (values in bold) are examples. The theoretical mean deceleration value can be any value below the track speed-dependent maximums.

2.2. Method

2.2.1. The Braking Behavior Measure (Dependent Variable)

The driving behavior is operationalized in one value for each red aspect approach, namely the mDtSPAD measure. At the start of this PhD, ProRail had already developed a proactive safety measure called Time-to-SPAD (TtSPAD) in cooperation with Dutch Railways (NS). During my PhD research, we developed a new safety measure called Deceleration-to-SPAD (DtSPAD) based on the previous measure. More information on the benefit of DtSPAD over TtSPAD can be found in section 7.2.3.

The formula to calculate the Deceleration-to-SPAD is

$$DtSPAD = \frac{0.5 * \text{current speed of the train}^2}{\text{distance to the red aspect}}$$

where speed is measured in meters per second and distance in meters.
The location and speed of the train are recorded by positioning sensors which are present on the trains. A deceleration-to-SPAD value can be calculated for every entry of speed and location supplied by the positioning sensor while the train is approaching a signal showing a red aspect. The DtSPAD indicates the deceleration rate the train needs to maintain to be able to stop exactly at the red signal. The maximum value of these is the mDtSPAD.

The DtSPAD calculation can start after a train has passed a signal showing yellow caused by a red aspect that is ahead (as start of a situation with a potential for an incident) and stops when the train is no longer approaching a signal showing either yellow or red. In these cases, the train driver has received authority to move on further, marking the end of a situation with potential for the incident.

In **Figure 16** the relationship between DtSPAD and actual deceleration is visible. The DtSPAD increases during an approach if the actual deceleration is lower than the DtSPAD value, and the DtSPAD decreases again if the actual deceleration is higher than the DtSPAD value.

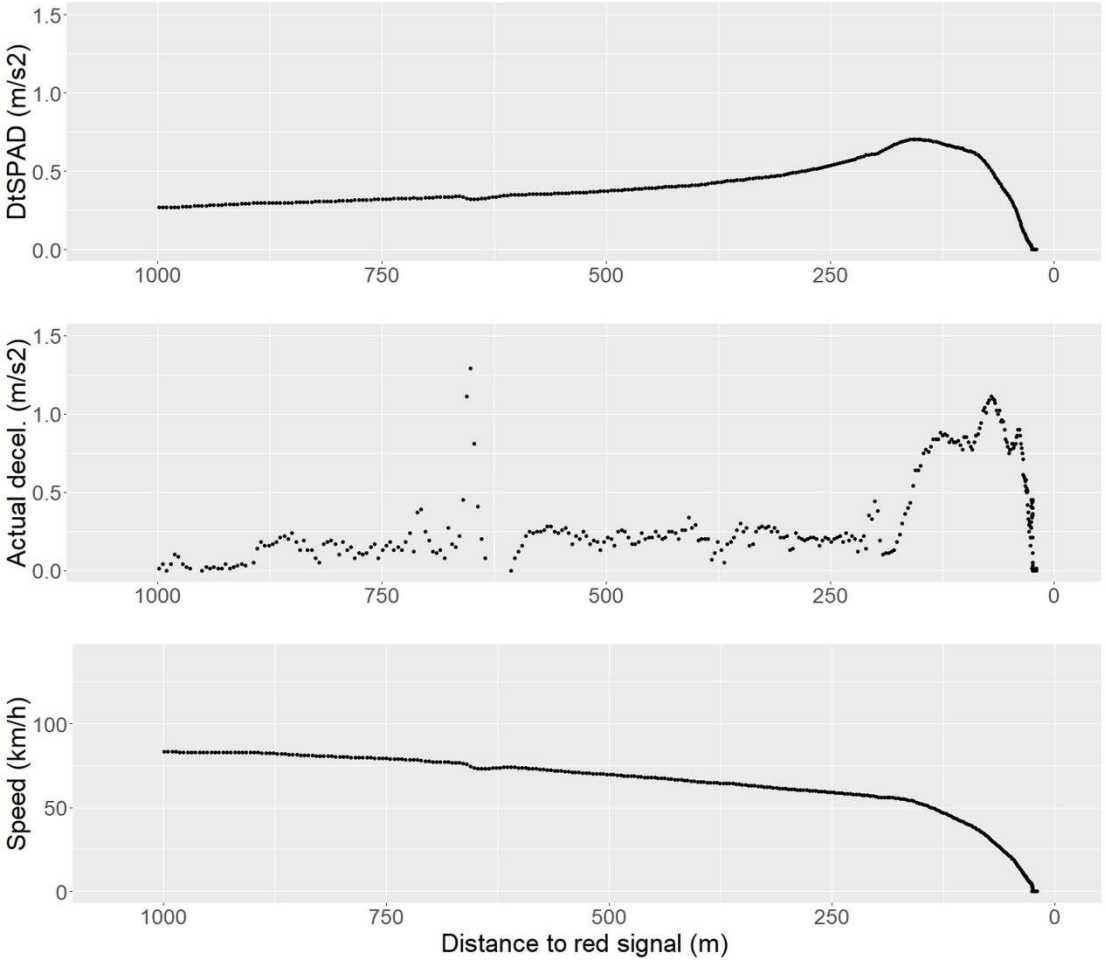


Figure 16. The risk indicator deceleration to signal passed at danger (DtSPAD), clarified using three different cross sections of the same red aspect approach. (top) Risk indicator over distance in meters (m). The DtSPAD reaches its maximum value at 152 m before the red aspect. (middle) Actual deceleration in meters per squared second (m/s^2) over distance. The DtSPAD declines once the actual deceleration is higher than the DtSPAD. (bottom) Speed in kilometer per hour (km/h) over distance graph.

The train's speed and position is needed to calculate the mDtSPAD. For this study, the information was gathered from Dutch Railways (NS) trains that have Orbit. Orbit is an auditory SPAD warning system. For this system to work, both the train's speed and position are registered, among other data. This data is logged multiple times per second from the moment the train is within 1000 m of a red aspect. Frequent logging (more than once per second) made this data source the most suitable. Automatic signals, which cannot be influenced by traffic controllers, are not monitored by the Orbit system due to technical limitations.

In our study we were interested in changes in behavior leading to higher mDtSPAD values. There is no absolute criterion for what constitutes a high DtSPAD value. In this study, 0.5 m/s^2 was chosen as a criterion for two reasons:

1. In previous initial analyses with similar data, the mDtSPAD followed a roughly normal distribution. The value of 0.5 m/s^2 was in the right tail of that distribution.

2. The Orbit warning system can alter the behavior of the train driver, and thus its mDtSPAD value, if the SPAD alarm sounds. For approaches where the Orbit alarm sounded, the mDtSPAD might have been higher if no warning system had been in place. In previous research it was noted that during most of the relevant approaches the alarm did not sound for DtSPAD values below 0.5 m/s^2. Unfortunately, the warning does not sound at a specific DtSPAD value. The algorithm for the warning system is based on other indicators that are not suitable for the current study.

Nineteen months of train data were analyzed, starting from 20 August 2018. On this date, approximately 50% of the trains of the operator NS had been equipped with Orbit (±300 trains). More trains were equipped with Orbit following this date, and their data were included as well. All were passenger trains with a brake power of up to 1.0 to 1.4 m/s^2. The train drivers were from the Dutch operator NS. The NS employs over 3000 train drivers and has 28 places of employment where train drivers start and end their shifts [65,66].

The Orbit system employs a quality filter to the GPS data. The warning system is temporarily shut down when the GPS quality becomes too low. In this study, we only used the data when the warning system was active. We also only included approaches where the time between two loggings was always below three seconds.

2.2.2. Inclusion Criteria

Braking behavior was calculated for the approaches falling within the hypothesis criteria and when:

- For speed: The track speed was higher than 80 km/h according to permanent traffic signs.
- For speed: The train did not pass a yellow aspect before the red aspect approach as part of a previous red aspect approach. Previous yellow aspects would have already resulted in lower train speed.
- For speed: The train was driving before passing the yellow aspect instead of departing from a station.
- For exposure: The red aspect remained red until standstill of the train or until the train was within 123 m of the red aspect. At 123 m, the train can still have a high value on our risk indicator at a speed of 40 km/h. This is the speed train drivers are instructed to decelerate to after having passed a yellow aspect to be able to stop for the red aspect.
- For other factors: The red aspect was not at a scheduled stop location. These approaches were excluded because the train driver would need to stop at these locations regardless of the aspect color.
- For other factors: The speed at which mDtSPAD was recorded was higher than 10 km/h.

2.2.3. Measures of Variation (Independent Variables)

The two independent variables were the theoretical mean deceleration and the frequency of yellow in last 14 days for this train series. Trains have the same train series when they are scheduled to drive the same route with the same stops. The theoretical mean deceleration (m/s^2) was calculated via 0.5 × track speed $(m/s)^2$/distance between signals (m). The frequency was calculated by counting the number of times the same train series passed the "yellow signal" in the last 14 days with a yellow+number aspect. Data from the Dutch infrastructure manager ProRail were used to calculate the frequency so that all train approaches could be used, not just those of trains with Orbit.

2.2.4. Tests Overview

An approach can be influenced by different effects. To deal with this overlap, the following tests were performed:

- To test the theoretical mean deceleration effect, approaches were selected where only the theoretical mean deceleration was a factor (exclusion of yellow entrance or yellow speed restriction; n = 3478 red aspect approaches).
- To test the yellow+number entrance effect, locations with speed restrictions were included if these speed restrictions had the same aspect. Three types of tests were done. The first test used all the approaches (n = 3429 red aspect approaches). The second test used approaches within a specific theoretical mean deceleration range (n = 2021 red aspect approaches for a high theoretical mean deceleration range and n = 1287 for a low theoretical mean deceleration range). The third test used approaches towards one specific signal. Only one signal was eligible as it had a sufficiently large number of approaches across different frequencies of entrance at yellow+number (n = 415 red aspect approaches).
- To test the speed restriction effect, approaches were selected where there were speed restrictions via yellow+number and a specific theoretical mean deceleration range. Locations with entrance at yellow were excluded (n = 509 red aspect approaches).
- To test the yellow–yellow–red effect, all yellow–yellow–red locations were included where there was no yellow+number speed restriction or yellow+number station entrance (n = 20 red aspect approaches).

See **Table 1** for a visual overview of the tested hypotheses and sample sizes.

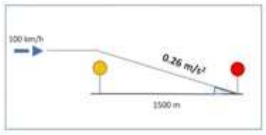
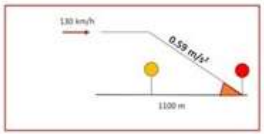
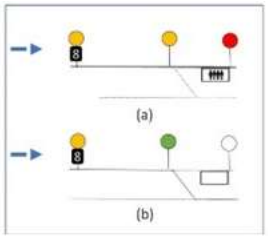
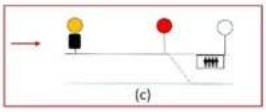
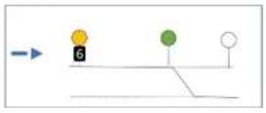
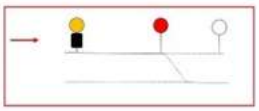
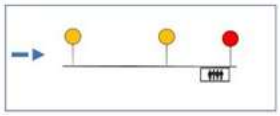
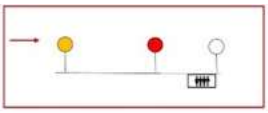
	Scenario with incidental learning opportunity	Scenario with possibility of error due to incidental learning	# Red aspect approaches (RAAs) tested
Theoretical mean deceleration effect			n = 3478 RAAs
Yellow+number entrance effect			<u>All the approaches:</u> n = 3429 RAAs <u>Specific theoretical mean deceleration range:</u> n = 2021 RAAs within the high theoretical mean deceleration range and n = 1287 within the low range <u>One specific signal:</u> n = 415 RAAs
Speed restriction effect *			n = 509 RAAs. * Insufficient data across different frequencies in previous 14 days to examine this effect separately
Yellow-yellow-red effect **			n = 20 RAAs. ** Insufficient data to examine this effect separately

Table 1. Overview per hypothesis of the performed tests and sample sizes. Displayed here as a visual reminder of the meaning of each hypothesis.

2.2.5. Statistical Analysis

To test the relation between the binary dependent variable and the (ratio) independent variables, a logistic regression analysis was considered. It was however not possible to perform a logistic regression analysis because the data did not fit the required assumptions of linearity of independent variables and log odds as shown by performing the assumption checks in the statistical software program SPSS.

Since the checks showed that there was no continuously increasing effect, we wanted to understand the actual shape of the relation. In order to investigate this shape, we considered a variation on piecewise regression. In piecewise regression, more than one line is fitted to the data. Multiple points in the independent variable can be chosen to split the data. These points of separation are called knots. Choosing the number of knots and their location is however very difficult. To refrain from using subjective input we decided to split the data evenly five ways. The first split was in half. The second split was in three segments, the third in four segments, the fourth in five segments, and the fifth in six segments.

The different splits lead to differences in under- and overfitting and in sample size per segment. Most importantly, insight is provided on the shape of the curve, which can be difficult with a binary dependent variable. The effect of knot selection is also shown. If the pattern remains the same across splits this is evidence for an effect.

The *p*-value[4] was calculated per segment by comparing the observed number of high mDtSPAD values with the number of high mDtSPAD values that is expected for the segment under the H0 assumption that there is no difference between segments (with 'high' being a mDtSPAD>0.5 m/s^2). The analyses were run in R, version 3.6.2. No additional packages were used for the analyses. The R Code is provided in appendix A. The steps are clarified with an example in **Figure 17**.

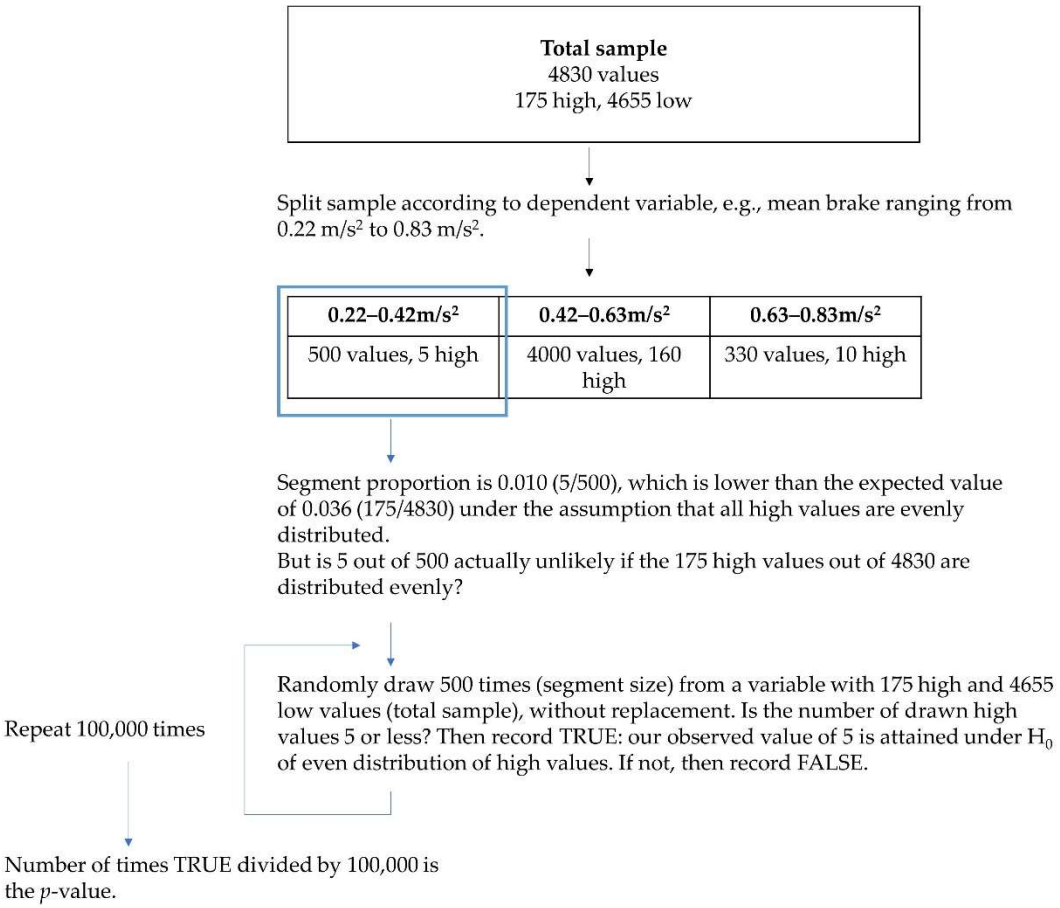


Figure 17. The p-value is calculated per segment by comparing the observed number of high values with the number expected if the high values are distributed evenly.

In the Results section the exact *p*-values were recorded when they were below 0.05, and were listed as $p < 0.001$ when they were below 0.001. *p*-values above 0.05 were recorded as non-significant (N.S.).

[4] In statistics, the p-value indicates the probability that the results can occur under the null hypothesis that there is no difference. A low p-value implies that it is unlikely that there is no difference and that it is more likely that the result is measured because of an actual effect.

2.2.6. Signal Effects

It is possible that there are signals which have many approaches with high mDtSPAD values. If the results are fully attributable to one or a few signals, the results are less likely to be caused by the investigated variable, i.e. incidental learning. To check whether the results were not fully attributable to one or a few signals, signals with more than three high values were identified. These signals are listed in the tables in the Results section and have been used to interpret the results.

2.3. Results

2.3.1. Theoretical mean deceleration Effect

If incidental learning occurs, we expect a correlation between the percentage of high values and the theoretical mean deceleration, with higher percentages for higher theoretical mean decelerations. Table 2 shows the results with the mean rate on the *x*-axis. Significant results were found for four out of five splits (i.e., the rows of the table). In general, the expected pattern was seen, with high percentages for higher theoretical mean decelerations (See **Table 2**).

1.00% 19 of 1903 p < .001			3.49% 55 of 1575 p < .001		
1.23% 5 of 408 N.S.		2.19% 64 of 2926 N.S.		3.47% 5 of 144 N.S.	
1.36% 3 of 220 N.S.		0.95% 16 of 1683 p < .001	3.38% 50 of 1481 p < .001		5.32%* 5 of 94 p = .048
1.66% 3 of 181 N.S.	1.12% 12 of 1071 p = .003	2.56% 53 of 2068 p = .020		1.20% 1 of 83 N.S.	6.67%* 5 of 75 p = .020
1.82% 3 of 165 N.S.	0.82% 2 of 243 N.S.	0.94% 14 of 1495 p < .001	3.49% 50 of 1431 p < .001	3.17% 4 of 126 N.S.	5.56% 1 of 18 N.S.
0.22-0.32 m/s²	0.32-0.42 m/s²	0.42-0.53 m/s²	0.53-0.63 m/s²	0.63-0.73 m/s²	0.73-0.83 m/s²
Signal A - 4 high values - 13.79% [theoretical mean deceleration = 0.57 m/s², speed = 130 km/u] *Signal B* - 4 high values - 7.69% [theoretical mean deceleration = 0.73 m/s², speed = 160 km/u] *Signal C* - 9 high values - 6.82% [theoretical mean deceleration = 0.56 m/s², speed = 130 km/u]					

Table 2. Theoretical mean deceleration effect analysis. This table shows the results when the theoretical mean deceleration is split five different ways. The percentages refer to the percentage of high values in a segment. The numbers directly below indicate the number of high values and the total. Orange indicates that the percentage is significantly higher than expected and blue indicates a percentage significantly lower than expected. White color indicates a non-significant result. The theoretical mean deceleration value ranged from 0.22 to 0.83 m/s². The mean percentage was 2.1%.

Can the effect be caused by an alternative explanation of signal effects? *Signal B* is almost solely responsible for the significant cells on the far right, annotated with an asterisk, contributing four out of five high values. This signal has a track speed of 160 km/h[5].

Separate inspection of approaches with track speed of 160 km/h showed that the percentage around a theoretical mean deceleration value of 0.6 m/s^2 seemed lower than those for approaches with track speeds below 160 km/h and a theoretical mean deceleration value around 0.6 m/s^2. It might be the case that approaches at a track speed of 160 km/h are experienced differently.

Potentially, (a) this highest theoretical mean deceleration segment in fact shows less behavior change and the effect seen is all due to *Signal B* with other unknown factors; or (b) there is learning within 160 km/h where the theoretical mean deceleration value of 0.6 m/s^2 is experienced as "much space" and only 0.7 m/s^2 as a "short" distance; or (c) the speed difference is attributable to chance and the effect on the outermost right cells is caused by the theoretical mean deceleration effect and not by a signal effect.

The high percentage for the segment between 0.53 and 0.63 m/s^2 is not attributable to specific signal effects, since there are 50 high values and only 13 of these are caused by two signals that have high percentages.

[5] The theoretical mean deceleration of 0.6 m/s^2 is the highest theoretical mean deceleration value permitted for track speeds up to 140 km/h in the absence of an inclining slope, as required by ProRail for signal distances. Thus signals with values above 0.6 m/s^2 will also have a track speed above 140 km/h.

2.3.2. Yellow(+Number) Entrance at Station Effect

2.3.2.1. Yellow Entrance at Station Effect: Analysis 1

The effect of yellow entrance is expected to increase with the frequency of yellow+number aspects at that location in the previous 14 days (i.e. a positive correlation between the amount of learning than can have occurred and the effect size). **Table 3** shows that the significant results follow the expected pattern of increasing percentages. The non-significant outer right percentages are however surprising. Since the total number of approaches is almost 500 for the outer right cell in the second split from the top, this percentage is most likely non-significant because it is close to the mean and is not due low power. The effect on behavior thus seems to taper off, rather than showing the expected continuous increase.

3.99% 107 of 2682 p < .001			7.23% 54 of 747 p < .001		
3.77% 95 of 2521 p < .001		9.90% 41 of 414 p < .001		5.06% 25 of 494 N.S.	
3.74% 92 of 2459 p < .001		6.73% 15 of 223 N.S.	9.30% 41 of 441 p < .001		4.25% 13 of 306 N.S.
3.68% 89 of 2418 p < .001	6.36% 11 of 173 N.S.	6.90% 16 of 232 N.S.		9.67% 38 of 393 p < .001	3.29% 7 of 213 N.S.
3.63% 87 of 2399 p < .001	6.56% 8 of 112 N.S.	7.45% 12 of 161 N.S.	11.46% 29 of 253 p < .001	6.33% 21 of 332 N.S.	2.47% 4 of 162 N.S.
0-87 times	88-174 times	175-262 times	263-349 times	350-437 times	438-525 times

Signal D – 27 high values - 22.9% [freq range: 0-3, with 91% of all approaches at 0]
Signal E – 40 high values - 9.6% [freq range: 0-473, with 93% between 200-500]
Signal F – 19 high values - 7.9% [freq range: 0-517, with 69% between 50-200)
Signal G – 14 high values - 5.0% [freq range: 0-465, with 77% between 200-500]

Table 3. Entrance at yellow effect (analysis 1).This table shows the results when the frequency in the last 14 days is split five different ways. The percentages reflect the percentage of high values in the segment. The numbers directly below indicate the number of high values and the total. Orange indicates that the percentage is significantly higher than expected and blue indicates a percentage significantly lower than expected. White color indicates no significance. The frequency of yellow+number aspects in the previous 14 days ranged from 0 to 525. The mean percentage was 4.7%.

Can the effect be caused by an alternative explanation of signal effects? A surprisingly high percentage of 22.9% was found for *Signal D.* Upon inspection by randomly sampling some approaches, it was noted that the preceding signal often showed the aspect *yellow+8* as part of a *yellow+8–yellow–red* sequence. This red aspect was however not at a station stop, which is why these approaches were not added in the calculation of the frequency. Despite the presence of this signal in the outer left segments, these segments are still significant on the lower end. The possible signal effect of S*ignal D* therefore does not affect the interpretation of the pattern.

The other signals have a wide range in frequency which would cause any potential signal effect to be spread out. The signal percentages were not higher than the highest significant cell percentages, making it unlikely that the pattern was fully caused by signal effects.

4.3.2.2. Yellow+number Entrance at Station Effect: Analysis 2

The previous analysis contained approaches with different theoretical mean decelerations. We know there is a significant effect of theoretical mean deceleration. Therefore, the test was repeated using only approaches in the theoretical mean deceleration range of 0.5–0.6 m/s^2. This segment was chosen because it was significant in the theoretical mean deceleration analysis (and not potentially explained by a signal effect like the theoretical mean deceleration ranges above 0.6 m/s^2).

Additionally, signal D was removed from this subset because there seemed to be a frequent yellow+8 aspect at that specific location which was not measured in our current method for frequency calculation (see Section 3.2.1).

Table 4 shows that the significant results still followed the expected pattern of increasing percentages. There are in fact more significant values, despite a smaller number of approaches. The low percentages on the outer right are surprising. The pattern remains of an effect that tapers off or even has an inverted u-shape.

4.69% 61 of 1302 p = .009				7.37% 53 of 719 p = .009	
4.25% 49 of 1152 p = .002		10.65% 41 of 385 p < .001		4.96% 24 of 484 N.S.	
4.22% 46 of 1091 p = .002		7.11% 15 of 211 N.S.		9.86% 41 of 416 p < .001	3.96% 12 of 303 N.S.
4.09% 43 of 1052 p = .001	6.51% 11 of 169 N.S.	7.77% 16 of 206 N.S.		9.69% 37 of 382 p < .001	3.30% 7 of 212 N.S.
3.97% 41 of 1033 p < .001	6.72% 8 of 119 N.S.	8.00% 12 of 150 N.S.	12.34% 29 of 235 p < .001	6.19% 20 of 323 N.S.	2.48% 4 of 161 p = .042
0-87 times	88-174 times	175-262 times	263-349 times	350-437 times	438-525 times
Signal E – 40 high values - 9.64% [freq range: 0-473, with 93% between 200-500] Signal F – 19 high values - 7.88% [freq range: 0-517, with 69% between 50-200] Signal G – 14 high values - 5.00% [freq range: 0-465, with 77% between 200-500]					

Table 4. Entrance at yellow effect (analysis 2a). This table shows the results when the frequency in the last 14 days was split in five different ways for the subset: theoretical mean deceleration value 0.5–0.6 m/s^2, without signal D. Percentages reflect the percentage of high values in this segment. The numbers below indicate the number of high values and the total. Orange indicates that the percentage is significantly higher than expected and blue indicates a percentage significantly lower than expected. White color indicates no significance. The frequency of yellow+number aspects in the previous 14 days ranged from 0 to 525. The mean percentage was 5.6%.

The prior analysis was repeated for the subset with theoretical mean deceleration smaller than 0.5 m/s^2. None of the splits led to significant cells. There were however relatively few approaches with a high entrance at yellow frequency (See **Table 5**). This caused problems with statistical power, especially because the number of approaches was very low in the middle section, which showed the highest percentages in the previous analyses. It is unknown whether there was too little power, or whether the yellow entrance effect was only present in combination with a higher theoretical mean deceleration.

1.51% 19 of 1259 N.S.			0% 0 of 28 N.S.		
1.52% 19 of 1250 N.S.		0% 0 of 27 N.S.		0% 0 of 10 N.S.	
1.52% 19 of 1248 N.S.	0% 0 of 11 N.S.		0% 0 of 24 N.S.		0% 0 of 4 N.S.
1.52% 19 of 1247 N.S.	0% 0 of 3 N.S.	0% 0 of 25 N.S.		0% 0 of 11 N.S.	0% 0 of 1 N.S.
1.52% 19 of 1247 N.S.	0% 0 of 3 N.S.	0% 0 of 9 N.S.	0% 0 of 18 N.S.	0% 0 of 9 N.S.	0% 0 of 1 N.S.
0-83	84-167	168-251	252-335	336-419	420-504
-					

Table 5. Entrance at yellow effect (analysis 2b). This table shows the results when the frequency in the last 14 days is split five different ways for the subset: theoretical mean deceleration value <0.5. Percentages reflect the percentage of high values in this segment. The numbers below indicate the number of high values and the total. Orange indicates when the percentage is significantly higher than expected and blue indicates when a percentage is significantly lower than expected. White color indicates no significance. The frequency of yellow+number aspects in the previous 14 days ranged from 0 to 504. The mean percentage was 1.5%. There were no signals with over three high values and a percentage above 3.0%.

4.3.2.3. Yellow+number Entrance at Station Effect: Analysis 3

The final analysis for the entrance at yellow effect contains data from one signal as described in the method section. **Table 6** shows two significant results in the expected direction. Most approaches were concentrated around the frequency of 300, leading to many cells with relatively few total approaches. Although the number of significant cells is underwhelming, the pattern displayed by the percentages is in line with the previous results.

2.63% 1 of 38 N.S.			10.34% 39 of 377 N.S.		
3.57% 1 of 28 N.S.		8.57% 9 of 105 N.S.		10.64% 30 of 282 N.S.	
3.57% 1 of 28 N.S.	0% 0 of 10 N.S.		13.92% 27 of 194 p = .005		6.56% 12 of 183 p = .040
3.70% 1 of 27 N.S.	0% 0 of 2 N.S.	11.54% 6 of 52 N.S.		11.81% 28 of 237 N.S.	5.15% 5 of 97 N.S.
3.70% 1 of 27 N.S.	0% 0 of 1 N.S.	0% 0 of 10 N.S.	9.47% 9 of 95 N.S.	11.16% 25 of 224 N.S.	8.62% 5 of 58 N.S.
0-78 times	79-157 times	158-236 times	237-315 times	316-394 times	395-473 times
N.A.					

Table 6. Entrance at yellow effect (analysis 3).This table shows the results when the frequency in the last 14 days was split in five different ways for the subset: one signal with theoretical mean deceleration value of 0.54 m/s^2. Percentages reflect the percentage of high values in this segment. The numbers below indicate the number of high values and the total. Orange indicates that the percentage is significantly higher than expected and blue indicates a percentage significantly lower than expected. White color indicates no significance. The frequency of yellow+number aspects in the previous 14 days ranged from 0 to 473. The mean percentage was 9.6%.

2.3.3. Speed Restriction Effect

Incidental learning was expected to influence driving behavior in locations where the signal aspect frequently was yellow+number due to speed restrictions. There were 509 red aspect approaches at locations with speed restrictions that were not at a yellow station entrance location. Unfortunately, 479 of those had a speed restriction frequency of 0 in the last 14 days. The remaining 30 approaches had a frequency between 1 and 15. There was thus insufficient data to examine this effect separately.

2.3.4. Yellow–Yellow–Red Effect

There were only 20 approaches that fell within the selection criteria. Many more approaches would have been present if approaches had included where the red aspect was at the station stop during the "deviating approach". Unfortunately, looking at these planned stops creates many methodological issues, including the influence of the distance between the red aspect and the stopping location.

2.4. Discussion

Can incidental learning contribute to SPAD incidents? In this study we took a step towards answering that question by first checking whether there was evidence of a change in behavior as a result of incidental learning. Significant results were found in the expected direction.

Other factors can however also influence the results, like signal effects. Deceleration behavior can be different for certain signals, for example because signal approaches differ in track speed, signal distance, and (early) signal visibility. The "entrance at yellow" effect was however also seen within one specific signal. That result cannot be influenced by any static signal effects. Other factors such as weather effects can play a role but these have no logical link with the frequency of yellow aspects in the previous 14 days and are therefore not likely confounding factors that created a spurious association.

The same result pattern that was seen for the one signal was also seen during the other "entrance at yellow" tests. The effect was therefore not only present for the one signal. Unfortunately, there was insufficient data to test whether the effect was also present for signals with a lower mean rate. It is therefore not yet known whether the "entrance at yellow" effect is always present, or only for those approaches with a higher mean rate.

It is possible that the approaches with a lower mean rate provide more time for the driver to correct his or her deceleration behavior before it shows up in our behavior measure. In theory, low mean rates might "buffer" against problematic situations. In the Netherlands, the trains are forced to decelerate at a minimal deceleration rate after passing the yellow aspect. This brings the speed down significantly for approaches with low theoretical mean deceleration rates in particular.

The shape of the effect was not entirely as expected for the entrance at yellow effect. The effect seemed to taper off as the entrance at yellow frequency reached very high values. Given the high frequencies, these were approaches where the train series had entrance at yellow almost every time.

A potential explanation is that the extreme familiarity with the situation leads to a heightened awareness when something is different. This is comparable to coming to a friend's house occasionally and going there nearly every day. When visiting occasionally one will recognize the picture on their living room wall. One might not notice when they change the picture to a comparable one. However, when the individual visits nearly every day he/she is more likely to notice that they changed the picture despite minimal changes.

It is of course also possible that there is a hidden factor that happens to be more present for those entrances with the highest frequency of entrance at yellow. This is unlikely, because a similar pattern was seen when looking within one signal, but the possibility cannot be excluded. Further research is needed to see whether the pattern is indeed caused by this psychological effect or whether it was an artefact of our data.

During the "entrance at yellow" effect, incidental learning occurred because the approach was in the same location and with a similar cue (e.g., yellow+4 and yellow). We also obtained evidence of a theoretical mean deceleration effect. In these situations, the location is different, but the cue is identical (yellow aspect).

The pattern for mean rate was as expected, with higher mean rates leading to higher percentages. However, the high percentages at the highest mean rates were caused by one signal and could thus be the result of a signal effect. Even if this is the case, the pattern remains for the low to medium–high mean rates.

It would however be jumping to conclusions to say that this pattern was definitely caused by incidental learning. It could be a conscious choice to always decelerate at for example 0.4 m/s^2, which would lead to a mDtSPAD above 0.5 m/s^2 for approaches with a mean rate above 0.4 m/s^2 and to low mDtSPAD values for approaches with a mean rate below 0.4 m/s^2.

2.4.1. Limitations and Future Research

In future research, additional factors could be included. One identified factor was the presence of a frequent yellow+number aspect caused by a red aspect that was not at a station stop. While the timetable is designed to avoid this kind of approach frequently in the same place, it is possible for this to occur. Additional involved factors could be line of sight, with early visibility as a protective factor.

An extra finding was the identification of signals with high percentages. It is clear that there are behavior-influencing factors that are currently out of scope and unknown. Whilst they did not interfere with the conclusions of this research, it would be an interesting avenue to discover what causes these differences between signals.

A limitation of our research was that the exposure frequency was calculated by train series and not by train driver. Since learning takes place in the mind of an individual, it would have been preferable to measure how often the train driver had previously experienced similar situations.

Information about the train driver was not disclosed for privacy reasons. The same train series was considered the next best alternative under the assumption that a train driver often drives the same train series.

Another possibility was to simply calculate how often any train was exposed to yellow aspects at the relevant location. We however assumed that train drivers link their experiences with the infrastructure to the train series they are in, since their driving experience is influenced by the present train series. A train driver might for example drive from Utrecht to Amsterdam, as many trains do, but the train series he is in determines which stations he has to stop at, what his timetable looks like, and the continuation of his journey.

Despite this limitation, the research was still possible because it focuses on relative changes. When a train series has an entrance at yellow frequency of 200 over the past 14 days, the specific train driver probably does not experience a yellow entrance in that location all 200 times. However, the train driver is likely to have experienced a greater number of entrances in yellow than in those cases where the frequency was only 100.

The Netherlands has 28 work locations for drivers, with each location having certain work packages, including some variation in routes but also repetition of routes by the same drivers [66–68]. Additionally, the authors analyzed Dutch SPAD reports and frequently noticed train driver statements such as “usually in this location there is aspect xyz”, further supporting the notion that the Dutch train drivers indeed drive the same routes repeatedly.

Nonetheless, the research would be improved by replication using driver data. This would also give more insight into how often an employee needs to be exposed to a certain situation for incidental learning to occur. Another related avenue for future research could be individual differences in incidental learning.

2.4.2. Answering the Question and Using the Answer

Our results indicate changes in train driver behavior when employees have previously been exposed to different behavior requirements in the same location with a similar yellow aspect. The results are in line with our expectations of incidental learning. Using data of actual everyday behavior, we identified a shift in braking behavior in the direction of a lower safety margin. We thus found evidence for the notion that incidental learning impacts employee behavior and thereby safety margins.

It is possible that the effects of incidental learning results in SPADs in certain situations. Further research can test whether the effects of incidental learning are indeed also visible using data of actual SPADs. A commonly known disadvantage of using incident data for quantitative analysis is that there is usually a small amount of data since there are relatively few (large) incidents.

This is especially the case in the Netherlands when looking at nuanced causes. There are for example multiple SPADs with aspect sequence green–yellow–red, but fewer with that specific aspect sequence and entrance at yellow. There are even fewer incidents within that segment with various frequencies of entrance at yellow (29 SPADs over 6 years as measured during the follow-up study described in chapter 3) . The results of this study, based on data of driving behavior, can be used to determine more exactly what aspects to investigate with incident data and in which manner. The follow-up study we performed using incident data is described in chapter 3.

The results of our study using deceleration behavior data can also be used as an input for decision-making on desired interventions. Crude measures, such as no longer using a specific signal aspect, are not necessary to eliminate certain behaviors or increase safety margins. We see that specific effects add up to create the locations with the highest percentages. **Figure 18** gives a simplified overview of how one signal approach can lead to different behavior depending on the theoretical mean deceleration, entrance at yellow frequency, and presence of speed restriction.

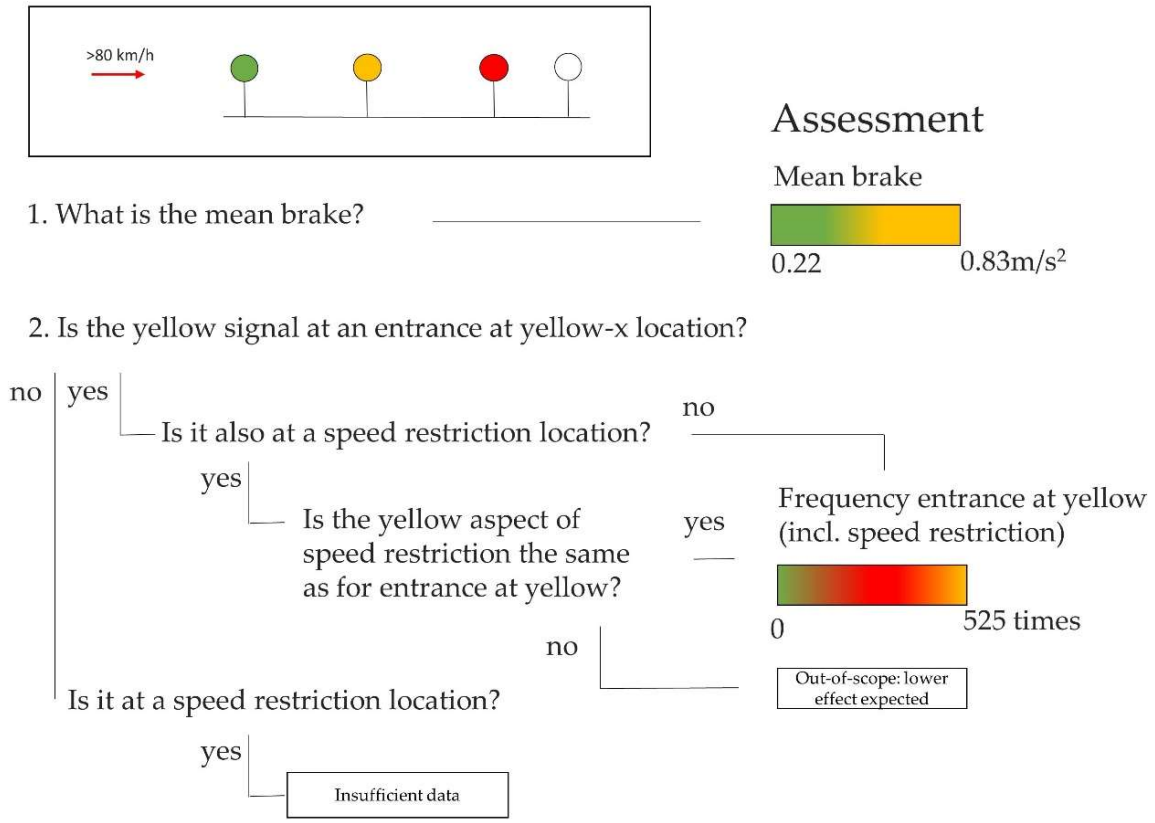


Figure 18. A simplified illustration that different factors need to be considered to predict differences in behavior.

In a general sense, organizations can reevaluate their task designs by taking the presence of incidental learning into account. Organizations often focus on making sure that the task design for a specific task helps the employee to perform the task successfully. This is important but does not address the whole story. To further improve task design, one should not only consider what the employee is exposed to during the execution of the specific task, but also what he or she has been exposed to during other moments of his shift. 'Yesterday' matters, especially if it is visually similar.

Chapter 3.

From error to incident and the window for correction in SPAD causation

Based on the article "What Employees Do Today Because of Their Experience Yesterday: Previous exposure to yellow+number aspects as a cause for SPAD incidents" by Julia Burggraaf, Jop Groeneweg, Simone Sillem and Pieter van Gelder

Chapter Summary

In chapter 2 it was shown that train driver deceleration behavior is influenced by exposure to less restrictive and visually similar signal aspects in the same location in the previous 14 days. Initial insufficient deceleration does not lead to a SPAD if the train driver adjusts the deceleration in time. In this chapter we will see that previous exposure to yellow+number aspects indeed only corresponds with a statistically significant increase in SPAD incidents if there is a small window for correction available to drivers (see **Figure 19**).

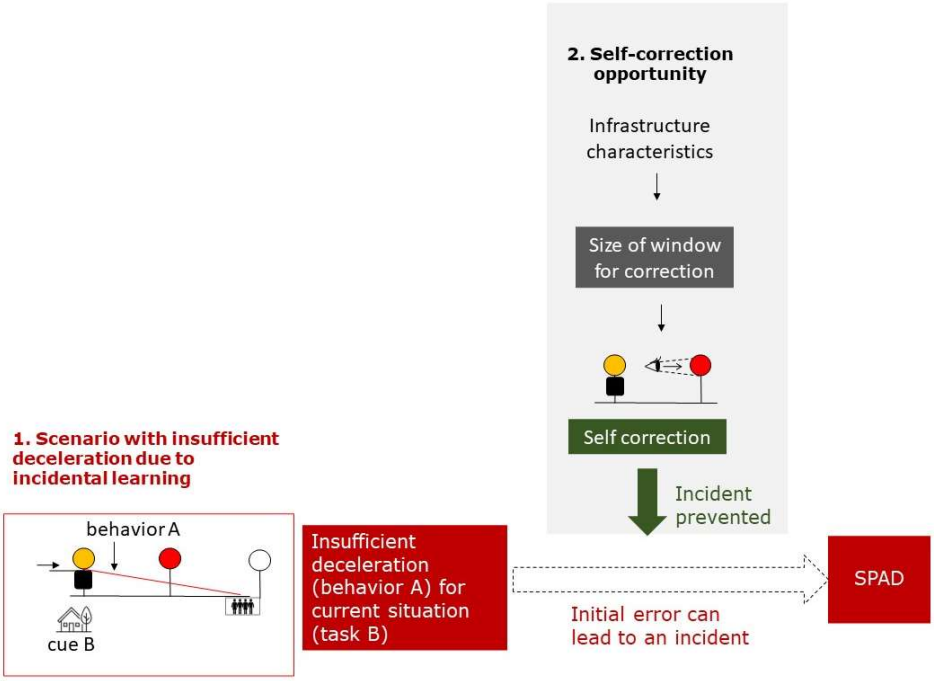


Figure 19. The error of insufficient deceleration can lead to a SPAD. Train drivers can however also self-correct the error upon perceiving the red aspect IF there is a large enough window for correction. The size of this window is influenced by the infrastructure design.

This figure is part of the larger figure presented in the dissertation summary (**Figure 3**) and the chapter 2 summary (**Figure 6**). This part of the larger figure is displayed here in the summary of chapter 3 to highlight that chapter 3 investigates the step from error to incident and the window for correction.

The size of the window in which train drivers can correct their initial insufficient declaration is influenced by infrastructure factors such as the permitted track speed and signal distance.

To test the role of incidental learning and the window for correction in accident causation, incident data was used. We used six years of SPAD incident data and red aspect approaches in the Netherlands for the analysis.

The results provide evidence for previous exposure as a cause for SPADs. Despite the significant and large effect of incidental learning on SPADs in combination with a small window for correction, there were only 13 SPADs where: A) the yellow+number frequency was larger than 150 and B) there was a small window for correction. This is

because there are relatively few RAA's with such high exposure frequencies ánd a small window for correction[6].

This chapter contains all the details on the specific infrastructure and timetable design situations for which an increased SPAD probability is identified so this data can be used to identify all locations in the Netherlands where this specific combination of circumstances occurs. Even more importantly, this knowledge can be used when designing new infrastructure to ensure that:

- these types of error-promoting infrastructure designs are not implemented
- or, if they must be used for other reasons, are implemented in combination with other safety barriers

This chapter thus provides concrete insights to be used within Rail, but even more importantly it provides insights in human-task interaction and the role in accident causation. When the task is not designed by taking the effect of incidental learning into account and the opportunity for self-correction, then the probability of an accident can increase.

[6] To put it into context: In six years, there were 141 thousand RAAs with a frequency over 150 and a small window for correction, while there were 13 million RAAs in the same period (both are RAA subsets of only RAAs without scheduled stop and without Orbit and passenger trains only). Thus only 1.07% of RAAs had a high exposure frequency ánd a small window for correction.

3.1 Introduction

In chapter 2 we saw that train driver behavior is indeed affected by the type of yellow aspect that was present in that location over the previous 14 days [69]. It is important to understand whether those measured changes in behavior can cause SPADs, in which situations and to what extent.

Balfe and Doyle analyzed a multi-SPAD signal in Ireland using the RSSB SPAD Hazard Checklist. The SPADs were preceded by a yellow aspect at a signal which usually showed a double-yellow aspect. The double-yellow aspect which shows in the standard situation indicates a red aspect at the end of the nearest platform, two signals further, whilst the yellow aspect (present during the SPADs) indicates that the next signal has a red aspect.

During all three SPAD events at the signal in front of the platform, the drivers reported that they understood the previous signal to be showing double yellow while it had in fact shown a single yellow aspect. Balfe and Doyle therefore identify the fact that the aspect is frequently double-yellow in that specific location as a potentially contributing factor to the SPADs [70]. The hypothesis is that a specific aspect (yellow in the above example) activates the behavior belonging to a different aspect (double-yellow in the above example). If the signal that showed the single yellow aspect during the SPAD event would have shown a green aspect on most previous approaches instead of the double-yellow aspect that was shown frequently, the SPADs might not have occurred.

In Dutch rail, double-yellow aspects are not used anymore, but other yellow aspect variations are common as we have seen in chapter 2. The section below explains why the initial error of insufficient deceleration may or may not lead to a SPAD. The train driver's "window for correction" is introduced, followed by the methods section, results, conclusion and discussion.

3.1.1 When wrong schema activation leads to a SPAD

In order for a SPAD to occur because of wrong schema activation in the brain, four elements should be present:

1. The train driver performs the deceleration behavior suitable in previous situations
2. There is a non-negligible difference in the current required behavior and the previous required behavior
3. The train driver does not correct his or her behavior in time
4. Technical (warning) systems do not intervene to correct the behavior

3.1.1.1. The train driver performs the past behavior

The previous behavior will be performed if it is activated sufficiently due to visual similarity and frequent past exposure (bottom-up) and the behavior is not prevented top-down by our "will" via our supervisory attentional system [71].

In chapter 2 it was shown that during the red aspect approach as part of the green-yellow-red sequence (last approach in **Figure 13**), the train driver behavior was indeed affected if the same train series had often passed a yellow+number aspect in the same location in the previous 14 days [69]. A higher frequency in the previous 14 days led to an increase in the change of behavior. This effect decreased after the frequency of a yellow+number aspect in the previous 14 days exceeded 400 times. This research indicated that train driver behavior is indeed affected by previous yellow+number aspects during yellow-red approaches.

3.1.1.2. There is a non-negligible difference in the required behavior and past behavior

Wrong schema activation will not cause a SPAD if a slight deceleration is sufficient during both previous and present approaches (for example, a deceleration of 0.31 m/s^2 is required in the previous scenario and 0.32 m/s^2 in the current approach). This minimal difference can occur when the track speed is low and/or the distance between the signals is large.

3.1.1.3. The train driver does not correct his/her behavior in time

A SPAD can be prevented if the train driver corrects his or her behavior in time. The theory of wrong schema activation predicts an initial insufficient deceleration after passing the yellow aspect. The train driver can realize the mistake and start to decelerate forcefully upon seeing the red aspect. Factors that increase or decrease the probability that a driver will be able to correct his or her behavior in time are expanded upon in section 3.1.2

3.1.1.4. Technical (warning) systems do not intervene to correct the behavior

A SPAD can also be prevented if other preventative interventions are present. The previously mentioned auditory warning system Orbit is designed to prevent SPADs by warning the train drivers with an auditory message when they approach a red aspect at a higher speed than desired. This warning system is not yet installed on all trains, nor operational for all signals. ERTMS can also provide preventative intervention but is not nationally implemented yet in the Netherlands. [11]

3.1.2 Investigating element three: The train driver does not correct his/her behavior in time

The four elements that need to be present for wrong schema activation to cause a SPAD have been described above. In chapter 2 we have seen that there is an effect of past yellow aspect exposure and thus the presence of element one. The risk of a SPAD is limited to those locations where there is a difference in past and current required behavior (element two), otherwise the past behavior is not erroneous, and where there are no technical (warning) systems that can intervene (element four). The remaining question is whether train drivers are able to correct their own behavior in time (element three) and prevent the wrong schema activation from actually causing a SPAD.

Whether the train driver will still be able to stop in front of the red aspect depends on the moment at which the train driver sees the red aspect, the braking distance as affected by the train's deceleration power and the track conditions, and the size of the window for correction. The concept of 'window for correction' will be illustrated by calculating the size of the window for correction for an infrastructure scenario example displayed in **Figure 20**.

The top of **Figure 20** shows a situation where the track speed is 130 km/h and the aspect at the first signal is often yellow+8, indicating a speed reduction to 80 km/h. The distance between the first and second signal is 1095 meters. In order for the train to drive at 80 km/h at the next signal, as is required, the train must decelerate continuously at 0.37 m/s^2.

At the bottom of **Figure 20** we have the hypothetical situation where the train passes a yellow aspect and decelerates at 0.37 m/s^2, as is commonly suitable in this location. However, since the aspect is yellow and not yellow+8, a continuous deceleration of 0.37 m/s^2 is insufficient to stop in front of the red aspect.

In this hypothetical example, the train driver sees the red aspect at 300 meters. He or she then attempts to correct his or her initial insufficient deceleration. If the train is able to continuously decelerate by at least 1.19 m/s^2 from that point onwards, then a SPAD can be prevented. Given that the emergency brake power of most passenger trains is around 1.2 m/s^2 and the train driver needs to initiate the emergency brake after perceiving the red aspect, this situation is one where there is a very small window for correction, assuming correct perception of the red aspect at 300 meters.

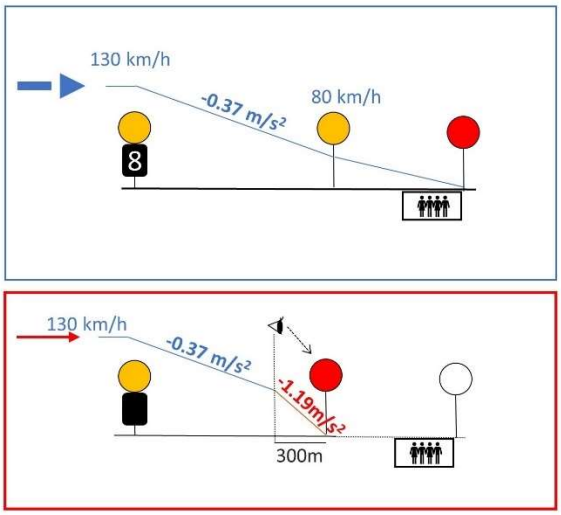


Figure 20. Hypothetical example where the standard situation with a deceleration of 0.37 m/s^2 is sufficient. If the train driver employs the same amount of deceleration during the green-yellow-red approach, then this insufficient deceleration can be corrected if, upon seeing the red aspect at 300 meters, the train can decelerate by at least 1.19 m/s^2 continuously.

It is unknown how often a train driver perceives a red aspect at 300 meters. The Dutch rail infrastructure manager ProRail has directives on the minimal distance at which a signal needs to be visible. The minimal visibility distance is 200 meters at track speeds below 80 km/h. At 100, 130 and 140 km/h the minimal visibility distances are respectively 250, 325 and 350 meters to maintain visibility for nine seconds per speed [72].

Even if the signal is theoretically visible, this does not necessarily mean that the train driver sees the red aspect and perceives it correctly. Seeing and perceiving the red aspect in time can also be influenced by a train driver's previous experience. Summerfield and Egner state in their review on visual cognition that visual detection and recognition are guided by one's prior knowledge of what is likely to occur [73].

Other factors can also influence early detection and recognition of the red aspect, like the visual conspicuity of the signal and aspect, weather conditions and situations where it is not immediately clear which signal belongs to one's track.

3.1.3 Research question and hypotheses

To further increase our knowledge of causes of SPADs, we investigate the following research question:

Does the frequency of a yellow+number aspect in the previous 14 days increase the probability of a SPAD if the window for correction is small?

This research question is answered for green-yellow-red aspect approaches where: the yellow aspect is in the same location as the yellow+number aspect, there is a difference between the previous required and current required deceleration behavior, and there are no technical systems present to correct the behavior.

We hypothesize that train drivers are able to correct their behavior and prevent a SPAD, if the infrastructure provides a large window for correction. However, when the window for correction is small, we hypothesize that wrong schema activation can contribute to SPAD causation.

Even though it is unknown when the train driver perceives the red aspect, it is possible to compare signal locations on the size of window for correction by comparing how much deceleration is necessary at a specific distance in front of the signal. Categorization as "small", "medium" and "large" windows for correction is discussed in section 3.2.4. This research will therefore also give more insight into which sizes of window for correction are large enough to provide the opportunity for self-correction.

3.2. Methods and materials

To answer the research question, we investigated whether there were more SPADs by passenger trains in the Netherlands between 2014-01-01 and 2019-12-31 than can be expected based on the number of red aspect approaches. This data included approaches by passenger trains of all operators. Six years of data were used to have as much data as possible in manageable quantities. Data from 2020 was not used because the timetable adjustments during the COVID-19 pandemic led to a large difference in the number of red aspect approaches.

For each SPAD and red aspect approach falling within the inclusion criteria, the frequency was calculated of a yellow+number aspect in the previous 14 days in the same location for the same train series and the size of the window for correction was calculated.

3.2.1 Data

Two main data sources were used: 1. SPAD incidents, and 2. Red aspect approaches. A list of SPAD incidents was provided by ProRail. Data of red aspect approaches was also provided by ProRail. When a train passes a signal, and the next signal is red at that point in time, it is recorded as a red aspect approach. The point in time at which a signal turns red or not-red is recorded for many of the signals on the Dutch rail network. For some signals, this point in time is not recorded, but can be deduced based on the moment that a train enters sections between signals. For some signals, this data is absent and therefore both SPADs and red aspect approaches at these locations were not included.

Whilst train kilometers are easier to obtain and therefore historically used more often, red aspect approaches are a better measure for the opportunity of SPAD occurrence [74,75]. The red aspect approach data also provides the additional details needed to test the hypotheses.

3.2.2 Types of SPADs

SPADs can occur for a multitude of reasons. This research focuses on driver error as a cause of SPADs and not on technical causes. The SPAD incident list was therefore filtered on SPADs that did not have a technical cause, or where the signal was put on red by the train traffic controller after the train had already passed the preceding signal. These SPADs could be excluded by only selecting the SPADs categorized as "non-technical – other".

The SPAD list did not include sufficient information in easily accessible format and was therefore enriched with data from the Red Aspect Approaches (RAA) dataset. The data was automatically matched based on date, train number and signal number. Only 47% of the selected SPADs could be matched with the RAA dataset. Upon inspection of the cases that were not matched, there were valid reasons why the match could not occur:

- The SPAD did not occur at a signal but at a sign
- The SPAD occurred when the train left the station whilst the departure signal was red, thus being a departure through red aspect and not a red aspect approach
- The SPAD did not occur with a passenger train but a road rail crane
- The SPAD occurred during shunting

The above situations are all outside the scope of this investigation. There were twelve SPADs that could not be matched because they were at a signal of which the red aspect time is not automatically recorded and could not be deduced. The twelve SPADs were at six different signals with six SPADs having occurred at one signal. These SPADs were within the scope, but not included in the analysis because they could not be matched. Including these SPADs manually was not an option, since the accompanying red aspect approaches should then also be added, which was not feasible.

3.2.3. Inclusion criteria for SPADs and red aspect approaches

The SPADs and red aspect approaches were included if they fit the criteria below.

Criteria to only select SPADs and red aspect approaches that are part of the hypothesis:

- The red aspect was part of a green-yellow-red aspect sequence
- The train was expected to approach with a speed higher than 80 km/h. The effect of exposure to past yellow+number aspects was only tested at speeds above 80 km/h because at low speeds the difference between past and required deceleration behavior tends to be small. Speed above 80 km/h is filtered in via 1) The track speed was higher than 80 km/h according to permanent traffic signs. 2) The train did not pass a yellow aspect before the red aspect approach as part of a previous red aspect approach. Previous yellow aspects would have already resulted in lower train speed. 3) The train was driving before passing the yellow aspect instead of departing from a station.

Criteria to avoid other effects influencing the analysis:

- The red aspect was not at a scheduled stop location. These approaches were excluded because the train driver would need to stop at these locations regardless of the aspect color.
- The above criteria related to speed excluded situations where two platforms were situated directly behind each other, thereby excluding situations where the red aspect was at the first platform whilst the train driver usually stops at the second platform.

- The yellow aspect was near a station stop as defined by being part of a red aspect approach to a scheduled stop at least once.
- The auditive warning system Orbit was not present or operating on the train. Since not all trains or all signals are protected by Orbit yet, it is still relevant to understand whether previous exposure to yellow+number aspects can contribute to SPADs when this warning is system is not present or operating correctly.
- For the statistical analysis which did not include the window for correction, only SPADs and red aspect approaches with a theoretical mean deceleration above 0.5 m/s^2 were included because the study described in chapter 2 showed a difference in driver behavior for these approaches [69]. For the statistical analysis which included the window for correction, the filter on theoretical mean deceleration is replaced by categorization based on the size of the window for correction.

3.2.4. Independent variables

The two independent variables were: A. the frequency of yellow+number aspect in the last 14 days for this train series and B. the window for correction. The frequency was calculated by counting the number of times the same train series passed the yellow+number signal in the last 14 days. Approaches are counted if (per train series) all have the same yellow+number aspect or if they do vary in yellow+number aspects with different numbers but the highest aspect frequency in the previous 14 days is above 100 and the other aspect frequencies are below 5.

The window for correction was calculated by taking into account what the yellow+number aspect was in the previous 14 days, what the distance was between the signals, and the permitted track speed. Since we do not know exactly what deceleration behavior the train driver usually employs, we calculate what the sufficient continuous deceleration is during the yellow+number approach. This is calculated via $\frac{v^2_{track\ speed} - v^2_{aspect\ speed\ limit}}{2 * distance\ between\ signals}$. We then calculate how fast the train should decelerate at 300 meters from the red aspect, if the train has been continuously decelerating with this deceleration level up until that point, via $\frac{v^2_{track\ speed} - sufficien\ \ continuous\ deceleration * 2 * (distance\ between\ signals - 300)}{2 * 300}$. We call this value "required deceleration upon 300 meters". In the example that was depicted in **Figure 20** with a track speed of 130 km/h (36.1 m/s), aspect speed of 80 km/h (22.2 m/s) and a signal distance of 1095 meters, the sufficient continuous deceleration is 0.37 m/s^2 and the required deceleration upon 300 meters is 1.19 m/s^2.

We categorized a required-deceleration-upon-300 meters value of less than 0.6 m/s^2 as a large window for correction, since this deceleration value is easily attained and very common. A value between 0.6 and 0.8 m/s^2 is categorized as a medium window for correction. A value above 0.8 m/s^2 is categorized as a small window for correction. This categorization is relative rather than absolute. It distinguishes locations with a larger window for correction from those with a smaller window for correction, rather than defining "large" and "small".

3.2.5. Analyses overview

Two analyses were run. The first one was performed to test whether the results in our study using deceleration behavior data could be replicated using incident data [69]. In this test, the window of correction was not included and the criterion for theoretical mean deceleration was included. The sample included 29 SPADs and 1,139,665 red aspect approaches.

The second analysis included the window for correction measure, which divided the SPADs and red aspect approaches into large, medium and small windows for correction. The window for correction could not be calculated for those approaches where there was no yellow+number aspect in the previous 14 days. These approaches were therefore not included in this test. A test was performed per window for correction to test the hypothesis that the relationship between frequency and probability of a SPAD only exists when there is a small window for correction. The samples included 0 SPADs and 777,510 red aspect approaches for the large window for correction, 3 SPADs and 319,533 red aspect approaches for the medium window and 17 SPADs and 54,462 red aspect approaches for the small window.

3.2.6. Statistical analysis

To test the effect of the yellow+number aspect frequency in the previous 14 days on the SPAD frequency, the same statistical method was used as used in chapter 2 to be able to see the exact shape of the relation between frequency and SPAD occurrence [69].

The bin sizes were based on the results of the study using driver behavior. Due to technical reasons and due to the large amount of data, it was necessary to select bins beforehand. It was therefore not possible to leave the frequency as an interval variable, which could potentially have been tested with a logistic regression. This was not considered a major issue since other testing is available and our previous study using driver behavior instead of SPADs showed assumption violations to perform the logistic regression anyway.

Since the relation between the yellow+number aspect frequency and train driver behavior showed a slightly inverted u-shape in our previous study and there were many approaches in the bins of lower frequencies, the following frequency bins were chosen: 0 times a yellow+number aspect in the last 14 days (only for the first analysis), 1–50 times, 51–150 times, 151–250 times, 251–350 times, 351–450 times, 451–550 times, >550 times.

The *p*-value was calculated per bin by comparing the observed number of SPADs with the number of SPADs that is expected for the bin under the H0 assumption that there was no difference between bins. The analyses were run in R, version 3.6.2. No additional packages were used for the analyses. The R Code is provided in Appendix B. The steps are clarified with an example in **Figure 21**.

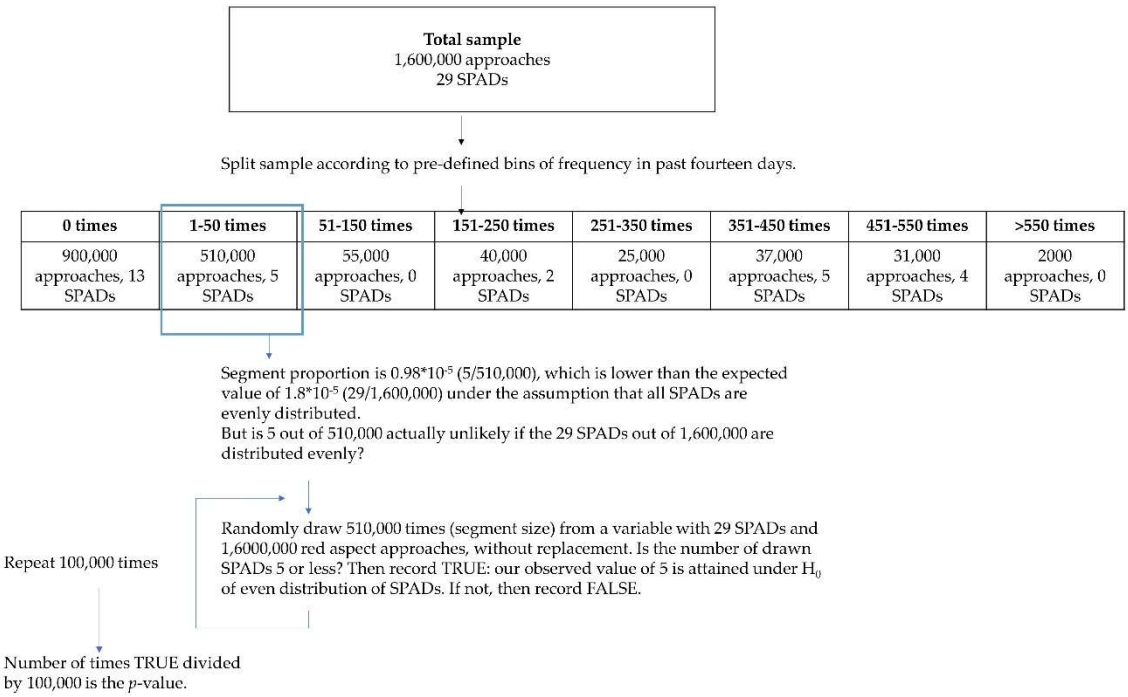


Figure 21. The p-value is calculated per segment by comparing the observed number of SPADs with the number of expected SPADs if the SPADs are distributed evenly.

In the Results section the exact *p*-values were recorded when they were below 0.05, and were listed as $p < 0.001$ when they were below 0.001. *p*-values above 0.05 were recorded as non-significant (N.S.).

Since there are relatively few SPADs, it is possible to have zero SPADs in a bin. If that bin is not significant, it is possible that: A. there is no difference in SPAD probability, B. the probability in that bin is lower but the result is not significant due to low power, or C. the probability in that bin is in fact higher but due to a low number of red aspect approaches in the given bin, no SPADs occurred.

If the average proportion, for example, is $1*10^{-5}$ SPAD per red aspect approach, then a bin with zero SPADs per 200,000 red aspect approaches is more likely to be non-significant due to option A or B since 1/200,000 is $0.5*10^{-5}$. A non-significant bin with zero SPADs per 30,000 red aspect approaches is more likely to be non-significant due to option C since 1/30,000 is $33*10^{-5}$. Non-significant bins with zero SPADs should therefore not be interpreted as low-risk.

Multiple two proportion z-tests or Fisher Exact tests were not used because the desired comparison was not to test whether two bins deviated, but whether a bin could be from the overall average, including that bin, violating the assumption of independence. A chi-square test for independence was considered as an overall test before the described test, but was not possible because of the violation of the assumption that the expected value per cell should be 5 or more in at least 80% of the cells, and no cell should have an expected value of less than one. The Fisher Exact test provided an out-of-workspace error in R due to the large number of variables and number of red aspect approaches. Other avenues to be able to perform the Fisher Exact test were not explored, since the statistical test for bin testing would already provide the desired answers, regardless of a preceding overall test.

Therefore, a simulation approach has again been used to test the above hypotheses.

3.3. Results

3.3.1 Yellow+number effect irrespective of window of correction

Table 7 shows that the number of SPADs is significantly higher for those approaches where the "yellow signal" was passed with a yellow+number aspect over 350 times and less than 550 times in the last 14 days. The bins with a frequency of 0 and 1–50 have the largest number of red aspect approaches and a SPAD percentage of respectively $2.00*10^{-3}$ and $1.55*10^{-3}$. The two significant bins have a SPAD percentage of $13.80*10^{-3}$ and $13.44*10^{-3}$, indicating not only a significant but also a large effect.

The absence of SPADs in the bin with frequency 251–350 is not in line with the hypothesis, but not surprising, given the relatively low number of red aspect approaches in this bin.

# times in previous 14 days	**0**	**1-50**	**51-150**	**151-250**	**251-350**	**351-450**	**451-550**	**>550**
# SPADs	13	5	0	2	0	5	4	0
# RAA	649,738	323,313	48,057	31,053	20,082	36,227	29,752	1,443
% * 10^{-3}	2.00	1.55	0	6.44	0	13.80	13.44	0
p-value	N.S.	N.S.	N.S.	N.S.	N.S.	0.002	0.006	N.S.

Table 7 SPAD percentage is significantly higher for a frequency of yellow+number aspects between 351 and 550 times in previous 14 days.

3.3.2 Yellow+number effect in combination with size of window of correction

Table 8 shows that the number of SPADs is significantly higher for those approaches where the "yellow signal" was passed with a yellow+number aspect over 350 times and less than 450 times in the previous 14 days and the window for correction was small.

Window for Correc-tion	**Frequency**	**1-50**	**51-150**	**151-250**	**251-350**	**351-450**	**451-550**	**>550**
Large (<0.6 m/s^2)	**#SPADs** **#RAA** **%* 10^{-3}** **p-value**	0 544,037 0 N.S.	0 103,675 0 N.S.	0 45,036 0 N.S.	0 38,098 0 N.S.	0 29,311 0 N.S.	0 17,002 0 N.S.	0 351 0 N.S.
Medium (0.6-0.8 m/s^2)	**#SPADs** **#RAA** **%* 10^{-3}** **p-value**	2 245,326 0.81 N.S.	0 30,958 0 N.S.	0 13,325 0 N.S.	0 7,996 0 N.S.	0 13,654 0 N.S.	1 8,063 12.40 N.S.	0 211 0 N.S.
Small (>0.8m/s^2)	**#SPADs** **#RAA** **%* 10^{-3}** **p-value**	4 345,537 1.16 0.001	0 61,080 0 N.S.	3 40,157 7.47 N.S.	1 24,695 4.05 N.S.	6 39,520 15.18 <0.001	3 34,545 8.68 N.S.	0 1,928 0 N.S.

Table 8 SPAD percentage is significantly higher for a frequency of yellow+number aspects between 351 and 450 times in previous 14 days and a small window for correction.

The SPAD percentage of $15.18*10^{-3}$ for the combination of frequency 351-450 and small window for correction is similar to the SPAD percentage in the previous analysis for the bin with the same 14-day frequency.

In the current sample, the bin with a frequency of zero in the past 14 days was not included, leading to a higher average SPAD percentage, potentially contributing to why the bin with frequency 451–550 is not significant in this analysis, apart from a lower, albeit still high, SPAD percentage. The higher average SPAD percentage can also explain why the SPAD percentage of $1.16*10^{-3}$ in the bin with a frequency of 1–50 is now significantly lower, whilst in the previous analysis it was also lower than average but not statistically significant.

No SPADs were measured in the category of approaches with a large window for correction. Only three out of twenty SPADs occurred in the category of a medium window for correction, but it should be noted that the number of red aspect approaches was low for this category and a 14-day frequency above 250.

3.4 Conclusions and discussion

The results indicate a significant and large effect of exposure in the previous 14 days in combination with a small window for correction. This evidence supports the hypothesis that incorrect schema activation is a significant contributor in SPAD causation if there is a small window for correction.

A limitation of this research was that the exposure frequency was calculated by train series and not by train driver. This limitation was also present in the previous study and expanded upon in section 2.4.1. Replication of this study using data with driver details would give more insight into how often an employee needs to be exposed to a certain situation for incorrect schema activation and execution to occur, which would give better guidelines for SPAD prevention.

In Dutch rail, locations with a higher probability of a SPAD can be identified if the following questions are all answered with "yes":

- Are there situations where yellow+number-yellow-red aspect sequences can also be yellow-red in the same location?
- Is the yellow+number aspect frequently present for that location?
- Is there a small window for correction, i.e., is there a large difference between the required deceleration during the yellow+number-yellow versus the yellow-red approach?[7]
- Are there no other SPAD prevention mechanisms present such as the auditory warning system?

Interventions to prevent SPADs can be aimed at preventing wrong schema activation, increasing the opportunity for self-correction and/or implementing intervention mechanisms.

To prevent the possibility of the wrong schema activation leading to an error, the infrastructure design can be improved via adjusted signal placement and/or track speed to make sure there are no locations where the yellow+number-yellow-red aspect sequence can also be yellow-red in the same location ánd there is a small window for correction.

When this is not possible for a given location, the probability of wrong schema activation can be reduced by a) decreasing the frequency of yellow+number aspects or b) increasing the dissimilarity between the yellow and yellow+number signal aspects.

To increase the opportunity for self-correction when wrong schema activation cannot be prevented, a) the window of correction can be increased, b) measures can be taken to increase the probability that the red aspect is perceived from afar (e.g. visibility and line of sight), and c) the braking power of the trains can be increased. Another option is

[7] ProRail has requirements for minimum distance between signals per speed and speed reduction. In the case of a zero percent slope and a track speed of 80 km/h, the size of the window for correction is 0.71 m/s^2 for the shortest possible distance in combination with a yellow+6 aspect. When the automatic train protection system ATB-EG is present in the tracks, the signal distance must be larger and the size of the window for correction is 0.62 for the shortest possible distance in combination with a yellow+6 aspect. The size of the window for correction is 0.51 m/s^2 or 0.46 m/s^2 for the shortest possible distance in combination with a yellow+4 aspect without and with ATB-EG presence respectively. At a zero percent slope and a track speed of 60 km/h, the size of the window for correction is 0.45 m/s^2 or 0.39 m/s^2 for the shortest possible distance in combination with a yellow+4 aspect without and with ATB-EG presence respectively. Thus, at a track speed of 80 km/h, the size of the window for correction is either large (<0.6 m/s^2) or medium (0.6-0.8 m/s^2) and at a track speed speeds of 60 km/h the size of the window for correction is always large (<0.6 m/s^2).

employing other SPAD prevention methods such as an auditory warning system or automatic intervention by technical systems.

The Dutch Railways started implementing the Orbit warning system in 2018. Initial results of in-company research projects within ProRail and NS have shown indications that Orbit has helped to prevent SPADs, although future research is needed to provide evidence that an auditory warning system such as Orbit can also prevent SPADs in which wrong schema activation occurred.

Nonetheless, understanding the causes of unsuitable deceleration behavior remains important even when Orbit is also a useful safety barrier in preventing these types of SPADs. Orbit simply does not cover all red aspect approaches because it is not implemented in all trains yet and is not designed to cover all signals. Technical failures are also possible. As is common within safety, we advocate the presence of multiple safety barriers, first of all by supporting the train driver to drive as desired by improving the infrastructure design.

Another reason that this research remains relevant despite technical advances is that it touches upon a larger topic, namely the need to take previous exposure into account for optimal task design:

- If (visually) similar situations often require one type of behavior and sometimes require different behavior, then the occurrence of an error should be considered.
- Employees can correct their own initial error if it is possible for them to perceive a clear signal (such as a red aspect) in time.

It is plausible that the incidental learning not only apply to the interaction between the train driver and signal aspects along the tracks, but also to the interaction between the train driver and on-board systems.

Chapter 4

Conclusions and discussion about incidental learning in Rail and other industries

4.1. Incidental learning as an important avenue for interventions and future research on human error

Incidental learning has been identified as one of the possible contributors to the occurrence of accidents via human error. The findings show the importance of not only evaluating task design by looking at the feasibility to perform a task at a given moment, but also how feasible it is to perform a task given what an employee is exposed to over time. Ideally, tasks are designed in such a way that incidental learning is unlikely to lead to errors. When this is not possible, situations can be identified where incidental learning is likely to lead to errors and measures can be implemented in order to prevent the errors from leading to an accident.

4.1.1. Relationship between frequency of exposure and error probability

Our research focused on a vital task within rail, namely the approach of a red aspect. Both the study using behavioral data and the study using accident data showed an effect of aspect exposure in the past fourteen days. The study using behavioral data showed an inverted-u shape between frequency and train driver behavior.

The largest effect was between frequencies of 260 to 350 yellow+number aspects in the past fourteen days. This inverted u-shape was seen for all relevant signals (both with and without theoretical mean deceleration filter) and for the one specific signal that was analyzed separately. The study using accident data showed a different relationship for all relevant signals with a high theoretical mean deceleration, namely an increasing and finally flattening relationship with the largest effect visible between frequencies of 350 to 550 yellow+number aspects.

There were differences in approaches between the study using behavioral data and the study using accident data. These differences are caused by differences in data availability and methods and by new insights and new research questions. An overview of the differences can be found in appendix C. The most important addition in the study using accident data was the calculation of the window for correction. The analysis that took window for correction into account did show the same relationship as was seen in the behavioral data study, namely an inverted u-shape relationship between frequency and SPADs.

As discussed in the discussion of chapter 2 on the behavioral study, this inverted u-shape could be caused by other factors than incidental learning. Theoretically, it is possible that the approaches with the highest frequencies also have a different characteristic. Many of the possible differences in type of approach are however unlikely to explain the findings, since the same shape was visible when looking at one specific signal.

One of the remaining options is that the approaches with the highest frequency in the past fourteen days were also accompanied by relatively more unplanned green-yellow-red approaches, creating more exposure to the high-risk situation and thereby moderating the effect of exposure to the yellow+number-yellow-red approaches. In that case, the underlying mechanism is still incidental learning but different exposure leads to different learning.

There is however no reason to structurally expect more unplanned green-yellow-red approaches at the highest frequency of yellow+number approaches in the past fourteen days in contrast to those approaches with a lower frequency. We also performed a brief check on a possible indication of a positive correlation between yellow+number frequency

and yellow frequency by randomly sampling eight time periods. For these eight time periods we looked at red aspect approaches towards the signal that was also separately analyzed in chapter 2. We looked at two different train series to reduce the possibility of a train series effect. Examining this data qualitatively gave no indication of a positive correlation between yellow+number frequency and yellow frequency and thus no support for this explanation. See appendix D for more information.

Alternatively, it is possible that the u-shape is inherent to incidental learning. Incidental learning could lead to the highest probability of wrong schema activation at moderate to high exposure, and less at very high exposure. As mentioned in the discussion of chapter 2, it might be easier to recognize that a situation is different than usual when there is a very high consistency in exposure. This idea is in line with the findings of Buttle and Raymond (2003) that high familiarity aids in change detection for face stimuli in comparison to merely familiar faces, even after an additional training period with the less familiar faces [76]. They use the term "superfamiliarity effect".

Future research could shed more light on how the frequency of exposure exactly relates to the probability of an error occurring, both during the task of driving towards a red aspect as during other tasks (in other industries).

4.1.2. Future research on circumstances

The current research gives rise to the question under which other circumstances incidental learning can also lead to errors, like other types of similarities during other tasks and within other industries. There are many factors which can potentially influence the probability of an error due to incidental learning. Further research on these factors will provide more insight into the exact mechanics and give more information for practical applicability in various settings.

4.1.2.1. Type of visual similarity

The findings indicate that the visual similarity between yellow and yellow+number signals is sufficient to lead to incidental learning. Additional research is needed to investigate what level of visual similarity is sufficient to lead to incidental learning in other scenarios or industries. In the current research, the main stimulus (the aspect) varied in terms of end-point deletion. In other words, there was either an additional number present or there was not. Other variations in visually similar stimuli are also possible, for example a mirror image reversal (e.g. 'ɜ' and 'ɛ') or mid-point deletion(e.g. '|' and '¦') [64]. Different types of similarity might lead to different probabilities of error due to incidental learning.

4.1.2.2. Other types of similarity

Another important open question is the role of the location. During the red aspect approaches, the two visually similar stimuli (the yellow aspects) were in the exact same location leading to: 1. a high visual similarity around the aspect and 2. a similarity in location. In other scenarios or industries, similar stimuli might not be in the same location. For example, a certain tool might be operated by turning to the left, whilst a different, but (visually) similar tool in a different location needs to be turned to the right. Or there might be a certain stacking rule in one warehouse, whilst a different warehouse might have a different stacking rule. What is the probability of an error occurring in those situations and for which levels of similarity? The role of auditory similarity can also be examined.

4.1.2.3. Variation in exposure

A component that was not investigated in the studies on train driver behavior, was the distribution between different types of exposure. In the rail studies, the independent variable was how often the yellow+number aspect was present in the past fourteen days, for a location where there was currently a yellow aspect. This gave no information about how often that specific signal was yellow (followed by a red aspect) in the past fourteen days.

In general, the yellow+number aspect as part of a scheduled approach is more frequently present at a given location than the unscheduled approach with a yellow. This gave rise to the labeling of yellow+number as the 'standard' approach and the 'yellow' approach as the 'deviating approach'. It was however not explicitly taken into account how often train drivers are exposed to the deviating approach in addition to the exposure to the standard approach.

When this information is taken into account in future research, it could also be relevant to include the sequence of exposure. A train driver might be exposed to the standard approach during 90 approaches and to the deviating approach during 10 approaching, but does it matter if the train driver is exposed to all 90 in a row before being exposed to a deviating approach versus for example 10 standard approaches followed by 1 deviating approach followed by a number of standard approaches and then a deviating approach again?

Especially when including the latter variable, data on train driver level rather than on train series level is recommended.

In the example of the warehouse, the effect of incidental learning might be different depending on whether the employee works 60 shifts in warehouse A and 20 shifts in warehouse B, versus 75 shifts in warehouse A, and 5 shifts in warehouse B. Additionally, the probability of an incidental learning effect might be higher during the first shift in warehouse B after multiple shifts in warehouse A, then on the fifth shift in warehouse B.

4.1.2.4. Opportunity for correction

The theory behind wrong schema activation predicted that the train driver makes an initial error of insufficient deceleration but can correct his or her behavior upon receiving new information, i.e. seeing the red aspect. The findings support the notion that the initial mistake does not necessarily lead to a certain form of tunnel vision, making the train driver "miss" the red aspect. This conclusion is based on the finding that the SPADs do not occur if there is considerable opportunity for correction.

Whilst this self-correction by the employee is never used as a sole safety barrier within railways, it is important to keep the room for correction in mind when adjusting the infrastructure and when performing research. Near miss situations because of self-correction might go unnoticed until a change in infrastructure that reduced the window for self-correction leads to SPADs.

The same heads-up applies to other tasks within railway in other industries: be aware that a large opportunity for correction might mask problems within the human-task interaction and that a reduction in opportunity for correction can be followed by an increase in the number of accidents if the rest of the system is not improved beforehand. The recommendation is always to 1. improve the human-task interaction to decrease the probability of an error occurring and increase the chance of processes running as intended and 2. having additional safety barriers in place that can prevent errors from leading to accidents which can include but should never solely be, sufficient opportunity for self-correction.

4.1.2.5. Value of stimuli

Signal aspects are vital for the train driver to perform his or her job well. The yellow aspect is especially critical, as it signals the presence of a red aspect at the next signal. The yellow and the yellow+number aspect are visually similar, but highly different in meaning. In future research, the meaning of the stimuli can be included. Is the probability of an error higher if the stimuli have no safety related meaning or does this have no effect?

4.1.2.6. Variation amongst employees

Another interesting avenue for future research is the variation among employees. Are certain people more prone to the effect than others and what does this depend on? Is there an influence of years of experience or of personality aspects? Are there also other factors that could make one especially prone to errors due to incidental learning, for example fatigue or beginning or end of the shift?

4.1.3. Problem with perception?

An alternative explanation to insufficient deceleration being caused by wrong schema activation could be that it is (simply) a matter of incorrect perception. There are however multiple reasons why incorrect perception is unlikely to be the cause.

Perception error of seeing what is not there?

If our investigated error corresponded with acting according to "yellow" when it was in fact "yellow+8", then the perception error would be more likely with the train driver not receiving visual input of the number eight or not processing it. This is however not the case for the tested scenarios. The investigated train driver error occurs when the train driver acts as if yellow+8 (or another yellow+number) was present, while in fact the aspect was yellow. If we would regard this in terms of perception, it means that the train driver perceived an "8" which was not present.

Clear visual input is not enough as demonstrated in the Stroop test

Even if we assume that the train driver could incorrectly perceive the number eight which was not there, then it still is not necessarily a perception problem. Research using the Stroop test (briefly mentioned in chapter 2) is a notable demonstration that even in the presence of clear instructions and clear visual input, the schema that is activated more strongly, wins.

During the Stroop test, participants have trouble naming the color in which a word is written if the letters spell a different color. The participant is for example asked to name the color of the blue letters that the word RED is written in. The correct answer is 'blue' in this example and the common error is saying 'red'. When asked to respond quickly, participants tend to make mistakes even though the visual stimuli is clearly visible and participants experience no problems with perception. They "know" the letters are in blue, but nonetheless respond by reading the word "red" out loud.

The effect is stronger when bilinguals participate in their dominant language than when they perform the task in their non-dominant language. This effect is attributed to differences in processing automaticity as a result of exposure and experience rather than to perception problems [77]. During the red aspect approach, the (visual) input is likely to activate multiple schema's within the brain of the train driver, where the strongest schema wins.

The complicated relationship between perception and action

What "perception" means and how it relates to action is also not a straightforward matter. We can see without visual perception and we can act before we consciously perceive or even without awareness of what was seen (see for example [78,79]).

There can be a difference between what we see and to what action it leads and what our brain reconstructs that we have perceived. The potential retro-active conscious naming of what the train driver has perceived, does not impact the action. The classical view is that sensory input leads to perception and then to action, but more recent theories show that action can actually influence early visual processing [80]. For more information see research on embodied cognition, such as [81,82].

Vision is not the only input for action

Finally it is important to realize that action is also activated by other modalities than just visual input. During the red aspect approach one could say: "the train driver responded as if having seen yellow+8", but one could also say: "the train driver responded as if being in this exact location".

In other scenarios, action might be triggered by visual ánd auditory ánd spatial input or even by previous action. Visual input does not solely determine our actions. Reducing the errors made by the train drivers to "they did not perceive the signal correctly" is an unrealistic simplification which, in worst case scenario, can lead persons to conclude that train drivers should simply "look closer" or "pay more attention", which is a counterproductive notion.

Reflections for future research: Considering findings from a road traffic study on change detection from an incidental learning perspective

The study described below is a simulator study into car driver behavior and driver response to the addition of a "no entry" sign at a road that was previously accessible. This type of variation in task design (addition of a "no entry" road sign) is far less common in everyday life than the presence of different aspects at a specific signal, whether on the road or within rail. This study is therefore more about behavioral response to environmental change than the incidental learning in the workplace where variation in stimuli and accompanying required different responses are part of an employee's every-day job. Nonetheless, the study results provide an interesting case to think about familiarity and exposure from a broader, incidental learning perspective rather than the more specific change detection perspective.

Martens (2018) used a simulator study to investigate whether people were less able to respond correctly to a "no entry" stop sign at an intersection when it was placed there after driving the same route eighteen times [83]. Of the participants that were confronted with the "no entry" sign on the nineteenth drive after eighteen identical drives without the sign, 30% entered the "no entry" road on the nineteenth. Interestingly, in the condition where the "no entry" sign was present on the first drive, 40% of the participants entered the "no entry" road. In the condition where the "no entry" sign was present after eighteen identical routes but with different scenery, 56% entered the "no entry" road on the nineteenth drive.

It should be noted that there were few participants, with only ten participants in the first two conditions and nine participants in the last condition. Nonetheless, it is interesting to note that multiple participants drove through the "no entry" sign even on the first drive. The authors attribute this to the road lay-out which was not in line with typical "no entry" locations. This explanation is supported by the fact that in conditions with (generic) warnings, there were zero to one participant driving through the "no entry" sign. In the conditions where there was a (generic) auditory warning before the intersection, there were no participants driving into the "no entry" road and in conditions where there was a sign with a generic warning about changed traffic situation before the intersection, there was only one participant driving into the "no entry" road for both the condition where the "no entry" sign was present during the second or the nineteenth drive.

The suggestion that road lay-out is more important than specific road exposure is interesting. In terms of familiarity, one could say that the condition with the "no entry" sign on the first drive had a certain level of exposure to the road lay-out based on everyday experience. The condition with same route but different scenery has a higher level of similar exposure and the condition with the same route ánd same scenery has the highest level of similar exposure. Here, we do see the same inverted-u shape: 40% error, 56% error, 30% error.

The exposure in this study consisted of only eighteen drives in a simulator during one afternoon and might not be representative of driving home every day on the same route. There were few participants and change detection was used rather than a stimulus that naturally varies as part of task design. However, this study does highlight some important factors to consider during research:

- Varying situations in terms of scenery whilst keeping required behavior identical
- Behavior triggers based on increasing exposure during the study
- Behavior triggers based on familiar road lay-out / task lay-out

4.2. Concrete application of the insights on incidental learning

We have seen that incidental learning can play a substantial role in causing human error in specific situations. We can build on the insights around incidental learning to provide recommendations for accident analysis, task design and research.

4.2.1. How to investigate incidental learning as part of an accident analysis

The main goal of analyzing accidents is to prevent future accidents by learning from the previous accidents. In order for effective learning, it is important to not only know what happened, but also why and what factors allowed for the accident to occur. There are multiple accident methods and models available and used worldwide. The methods differ in complexity and scope, ranging from 'asking 'why' five times' to drawing non-linear models which include factors outside the organization.

Part of all methods is identifying why a human error occurred (e.g. in the 'five-times-why' method) or what the context was that increased the probability of a human error occurring (e.g. Tripod, HFCAS) [84]. Some methods, like HFCAS, provide taxonomies of error contributing factors. These taxonomies provide guidance for the incident investigators [85]. Accident analysis is a complex cognitive task. It has been found to be susceptible to cognitive biases [86]. Providing lists of possible error contributing factors can help incident investigators to systematically evaluate whether they have considered multiple options.

The Dutch Safety Board created an evaluation checklist for accident analysis methods. The questions included: 'Does the method stimulate discussion between the team of investigators?', 'Can the method be taught easily?' 'Does the method stimulate investigators to identify a broad range of factors (minimizing tunnel vision)', 'does the method discourage pointing towards a culprit?' [84]

We believe that adding incidental learning as a factor to consider, would improve the accident analysis method on these aspects, regardless of which method is being used. For methods such as HFCAS which already have taxonomies, it is recommended to add incidental learning to the list of possible contributing factor.

4.2.1.1. Four incidental learning questions to ask during accident analysis

The generic recommendation for accident analysists is thus to consider whether an accident could have been caused by incidental learning. One can go about this by asking a set of questions. These questions are listed below with example answers given for an accident in rail based on the research in this dissertation, an aerospace example inspired by an actual aerospace accident in the Netherlands [87], an inland shipping example inspired by an actual accident in the Netherlands [88] and medical example inspired by an actual accident in the UK [89].

Please note that, although these examples are based on actual accidents, they are simplifications of the actual situations that occurred. Assumptions were also made about the processes within the respective industries which might not be entirely correct (e.g. how often certain tasks are performed) but were made for illustration purposes.

- **Q1: What was the task where a human error occurred? What was the behavior that was the human error?**

Rail example: The train driver approached a red aspect. The train driver decelerated insufficiently.

Aerospace example: The pilots positioned the plane on the runway for lift-off. The pilots positioned the plane incorrectly at the edge of the runway instead of the center.

Inland shipping example: A tanker sails towards a closed barrage using radar due to heavy fog. The captain sails the tanker into the closed barrage.

Medical example: A nurse administered intravenous medication to a patient after a pacemaker had been fitted. Administration of the penicillin was an error because the patient had a known penicillin allergy.

- **Q2: Is it a type of behavior which can be performed by the employee with relatively little attention / with high automation?** This is for example the case if it is commonly performed behavior. The behavior should not require looking up in a manual how to do it or asking someone else, nor should it require thinking pauses to consider how to perform the next (manual) action.

Rail example: Decelerating the train towards a signal is a key behavior for train drivers and very common. They are likely to be able to perform it with little attention/high automation. Conclusion: yes.

Aerospace example: Positioning the plane on the runway is a very common behavior. Pilots are likely to be able to perform the manual task of steering and decelerating and accelerating with little attention/high automation. Conclusion: yes.

Inland shipping example: Because of the fog, the captain had to sail based purely on radar which is very hard and requires a lot of attention. It is also far less common to sail for long periods in closed fog using radar than it is to be able to sail with visibility. It is unlikely that the captain could behave with little attention/high automation. Conclusion: no.

Medical example: The error lies in the choice to administer penicillin rather than in the manual steps of administering medication. Whilst this error thus does not typically revolve around behavior, one can say that the cognitive step from 'patient who has just had a pacemaker fitted' to 'I should administer penicillin' is one which can be performed with high automation in the sense that little explicit thinking effort is required. Conclusion: yes.

- **Q3: What did the employee see, hear, smell, feel or had just done at the moment of behavior execution? Was what they saw, heard, smelled, felt or had just done similar to what they see, hear, smell, feel or do in other situations where the performed behavior is in fact suitable and not an error?**

Rail example: The train driver saw a yellow signal at a specific location. In other situations, the train driver can see a similar signal aspect (yellow+8) in the exact same location. The lower amount of deceleration which was an error during the accident, is suitable behavior in the situation with the yellow+8 aspect. Conclusion: Yes.

Aerospace example: During the incident, it was dark outside and there were runway edge lights which were on. There was no taxi center line lighting. Visually, there was therefore one line of lights. This is visually similar to other situations where there is one line of lights at the center of the runway. Positioning the plane in front of the line of lights was an error during the accident, but is suitable behavior in the situations where there is only a taxi center line lighting. Conclusion: Yes.

Inland shipping example: The captain could not look outside due to heavy fog. It is possible that the visual cue of the radar is similar to other times when he sailed on radar in the same location, but heavy fog is not that common. Additionally, the barrage at the specific location is hardly ever open so the behavior of continuing to sail would be an error during both the accident and during the similar situation. There might be some elements within the radar that were similar to other situations where continuing to sail is suitable, but in this scenario that was not the case. Conclusion: No.

Medical example: Intravenous penicillin is the usual antibiotic used following a pacemaker being fitted. Insufficient details were provided on what the visual similarities were in the situation with this patient with the known allergy and the generic situation of other patients without a known allergy. However, there was no clear record of the allergy in the medical notes, so it is likely that there were visual similarities within the notes between the patient who could not have penicillin administered and those who can. The patient's allergy band was covered with a bandage for an intravenous drip, so there was no visual differentiation on the patient's body. Conclusion: Yes.

- **Q4: Is the similar situation in which the behavior is in fact suitable, commonly present per employee?** If there is data available on the frequency, then this is very useful, but otherwise more qualitative questions can be asked, like: at the very least, does it ever occur in a month?

Rail example: Some signals can have a different, but similar signal aspect at the same location, while others cannot. For those signals that can have different, but similar signal aspects, those that are near a station are more likely to have frequent yellow aspects than others. In our example, the signal that was yellow can also display the aspect yellow+8 and is likely to do so frequently because it is near a station stop. Conclusion: yes.

Aerospace example: Pilots commonly fly during the dark on runways which have only taxi center line lighting and therefore often see the one line of lights and have to often position the plane in front of it. Conclusion: yes.

Inland shipping example: Not relevant given the no at the previous question.

Medical example: Administering intravenous penicillin to patients who have had a pacemaker fitted is common behavior for nurses. Conclusion: yes.

- **Conclusion**

If question two, three and four are answered with 'yes', then incidental learning is likely to have played a role. If any question is answered with 'no', then incidental learning is unlikely to have played a role.

However, the most important function of above questions is to inspire incident investigators to consider whether there are similar situations whose previous exposure

could have increased the probability of an error due to similarities in how the situation looks/sounds/feels/tastes and the required performance of automated behavior.

As a final note, be aware that it is not fruitful to directly ask the employee involved in the incident whether he or she was influenced by incidental learning. As discussed previously, when schema's are present, one can perform actions with less attention. Attention is an important component of encoding memories which is why behavior performed with less attention is less likely to create a memory [90,91]. This is the reason many of us find ourselves rushing home to check whether we have turned off the stove, closed the window or locked the door, when we have already done so. The common behavior of turning off the stove or locking the door does not create a vivid memory in our minds.

4.2.2. How to investigate incidental learning as part of task evaluation and task design

Tasks are the specific activities of an employee which are needed to achieve functions and thereby goals [30]. Whilst there can be varying degrees of freedom in how an employee performs a task, there is usually a (physical) system that the employee needs to interact with and a set of procedures or rules on how the task should be performed.

Both the system and the procedures can be designed in multiple ways. Here we refer to the design of the system and the procedures involved in the execution of a task as the 'task design'. Ideally, a task is designed in such a way that the probability of an error is as low as possible. Alternatively, safety barriers can be incorporated in the task design to ensure that errors do not have significant undesirable consequences if the errors cannot be provided by improved task design.

When it comes to the evaluation of both existing and new tasks, involvement of (prospective) users is often advised. Usability testing, in which users interact with the system to identify design flaws overlooked by designers, can be a useful way to evaluate and consequently improve task design [30].

However, when it comes to the effects of incidental learning, these issues might be overlooked during usability testing because of relatively low error probability, insufficient previous exposure for learning to occur in the test setting and increased attention of the users in the artificial setting. Heuristic evaluation can also be performed, which refers to a systematic evaluation of the design to judge compliance with human factors guidelines and criteria [30]. Similar to the recommendation for accident analysis, we advise the inclusion of incidental learning as a factor to evaluate upon during heuristic evaluation.

4.2.2.1. Six incidental learning questions to ask during task evaluation

When a task design is evaluated on the effects of incidental learning, the following questions can be asked:

- Are there any tasks where human error can lead to an accident? For such a task:
- What is the cue for the employee to perform the behavior? What does he or she see or hear or has to do previously (or even smell or feel) that elicits the behavior? List 'cues A' and 'behavior A'
- Are cues A similar to the cues that are present in other situations where different behavior is required? List 'cues B' and 'behavior B'
- What happens if behavior B is performed during the task where behavior 'A' is required? Does this lead to a problematic error? If yes, continue answering the questions. If no, consider whether the design can be accepted.

- Is behavior B more commonly performed by an employee than behavior A? If yes, continue answering the questions. If no, there is less chance of a problem, although errors with high consequences should still be taken considered.
- Is there an opportunity for correction? Is the employee triggered by a system which detects the error or does the employee have to detect the error on his or her own? If the employee acts correctly upon detection of the error, will this still be in time to prevent the error from leading to an accident? Depending on the answer and the severity of the error consequences, consider improving the task design.

Above questions can be asked for the functioning of the system as designed. Another element to consider is that employees also learn from exposure to suboptimal system functioning. For example, frequent 'false alarms' by a system can lead to an increased probability of responding to the alarm by ignoring it or immediately clicking it away without further actions as is often experienced to be unproblematic but can lead to an accident in those situations where the alarm was in fact correct. Such situations can for example occur for auditive warning system but also for intervening systems such as automatic emergency brake.

4.2.2.2. Example of tasks where evaluation on incidental learning is useful

One specific type of tasks not yet mentioned in previous examples are those using digital interfaces and automation. Automation and the use of interfaces is becoming increasingly prevalent. When employees are exposed to differently designed interfaces which require different actions upon perceiving similar visual or auditive messages, then problems can occur. One type of error that can occur within using the same system is the 'mode error' where the user assumes that the machine is in one state when it is in fact in another [92]. Depending on which mode the system is in, the same action might be either suitable or an error.

Options to consider to improve the task design:

- Prevention: Make the cues that are present in situation A and B more dissimilar
- Prevention: Make behavior B impossible to execute in situation A
- Prevention: Make the execution of behavior B only possible in situation A after additional action steps
- Preventive safety barriers: Increase the opportunity for error correction by implementing systems which detect and warn early enough for a corrective action to not lead to an accident.
- Mitigating safety barriers: Anticipate uncorrected errors and add safety barriers which ensure that the error does not lead to damage, harm or injury.

Chapter 5

Research inspiration by analyzing variation: A Shewhartian view on process safety

Based on the article "Everything under control, check your variation! A Shewhartian view on process safety and other applications" by Julia Burggraaf, Frank Guldenmund and Jop Groeneweg.

Chapter summary

Data can be analyzed in many ways. We have found that examination of the variation in human behavior can help identify relevant human behavior questions for in depth research. This is especially the case if there is more variation in behavior among certain tasks while this is not expected based on the characteristics of the tasks.

If organizations are in control of their systems and processes, they will be highly predictable with little variation. Some level of variation is inevitable: even when someone attempts to do the same task twice in an identical way, the outcome will always be slightly different. From a process safety perspective, the relevant questions are: how much of that variation is acceptable and part of the normal accepted way of working and what are sources of unwanted variation? This approach was developed by Shewhart: he placed an emphasis on reducing variation within quality management. High variation, especially if the source is not known and the occurrence not understood, can be a mark of lack of control over organizational processes. Incidents in this view are not the result of special causes but are the extreme cases of normal operations with too much variation.

We advocate examining, understanding and reducing unwanted and avoidable variation to improve safety performance and use the amount of variation as indicator whether an organization is in control over their processes and, hence, its safety. The developments within technology, including sensors and data analysis, now make it possible to map the variation of processes and outcomes that are of interest to process safety. Once processes are measured, comparing variation between processes that are considered to be similar or considered to be different in variation, can be used to gain a deeper understanding of such variation. This Shewhartian view on process safety is clarified with a case that will illustrate the need to take on this approach to increase our understanding of human behavior and further reduce the number of incidents.

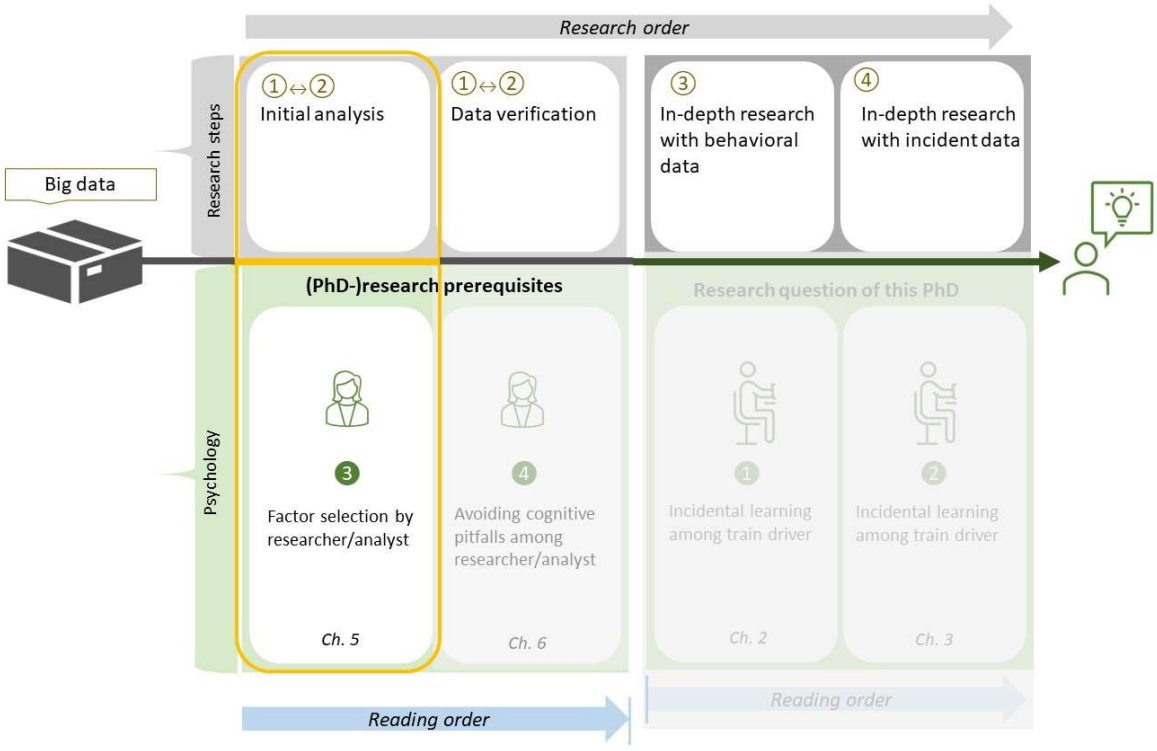


Figure 22. Location of chapter 5 with respect to research order and reading order. Chapters 2 to 4 discussed the research question of this PhD. Chapters 5 and 6 discuss the research prerequisites.

5.1. Introduction

The mathematician Jacob Bernouilli consulted Leibniz on a notion he had: Is it possible to calculate the probability of a man of twenty outliving a sixty-year-old man, using gravestone data? Leibniz replied that this might not be possible, as there is no limit to the sources of variation in nature, i.e. death causes, that will appear in the future [93]. In other words, next to the common causes of variation that will always be present in any process, new and unpredictable causes will appear, call them special causes of variation, which we will never be able to predict in the present, let alone control.

The correspondence between Bernouilli and Leibniz occurred in 1703. Fast-forward to the '30s of the last century, when production plant managers struggled with significant amounts of scrap, products not meeting specifications. Quality control at the time was limited to inspecting finished products, which gave little away of what went wrong and where. Walter A. Shewhart, born in 1891, addressed this issue within quality management by focusing on the causes of the variation in outcome. In 1931 he formulated as his third postulate: "Assignable causes of variation may be found and eliminated" ([94], p.14).

Shewhart also developed control charts which are relevant when it comes to monitoring an in-control-system over time. His control charts are famous, widely used and beneficial. These charts are used as an early warning system to see, over time, if a process is going out of control. These control charts take the distribution and variation of a process into account, based on the idea that values can vary due to the "natural" variation of a process (*common* variation), yet values become alarming when they are outside the expected values given their "natural" variation [94,95].

One of the big challenges organizations face with respect to process safety, is answering the question: Are we fully in control? The presence of incidents can indicate that the organization is (partly) out of control, depending of course on the accepted level of risk in the process. However, the absence of incidents does *not* equate to being in control. The field of safety is not the only field concerned with answering the question whether a process is controlled or not. Insights from the quality domain can be used to increase the understanding of why incidents occur and identify new ways of reducing them.

Central in Shewhart's approach is the idea that sources of variation (noise) must be identified and eliminated [94]. This idea can also be applied to process safety. In the view of Shewhart, accidents are "just" extreme outcomes of a system that allows for too much variation and is therefore not in control (Ibid.). This contrasts with the more traditional approach towards safety: causes of incidents must be identified and eliminated. Of course, the causes identified via the traditional approach can be sources of unwanted variation and are interesting targets for improvement. What is added in the approach of Shewhart is that the amount of variation in the normal way of operating should also be the target of an investigation.

Discussions within the safety domain on process safety indicators and early warning indicators show little mention of Shewhart's control charts and his focus on distributions and variation (e.g. [96,97]). Best wrote that in 1924, "Shewhart described the first control chart" in the Hawthorne factory, where in November of the same year, "a series of research projects began which came to be known as the Hawthorne studies. This body of work was central to the creation of the fields of the sociology, social psychology, and anthropology of the workplace." While both events occurred around the same time at the same place, these streams of ideas did not become entwined. Shewhart's notions

influenced engineering and production management, yet not the social science and organizational behavior that underly a lot of thinking within safety [98].

Shewhart's ideas of focusing on variation have been around for a long time. Despite having proven their merit within quality management, they have not caught on within safety, so why haul them up now?

The answer is twofold: firstly, the traditional approach within safety have led to many improvements in domains like aviation, refineries, and railways, but organizations are now faced with a residual number of accidents that prove to be elusive in their causes, when investigated via the traditional approach. Secondly, developments within data theory and data technology have finally advanced sufficiently to make it practically feasible to implement this notion.

Taking into account the distribution and the variation of pertinent output and process parameters (or parameters deemed pertinent), requires collecting, storing and analyzing data of all situations with the potential for an incident, and not just those situations in which the potential actually turned into an incident. Previously, data was often stored temporarily and overwritten after a day, week or month, because the amount of data was simply too much to store and analyze. This still happens, but less often because of costs or technological constraints, but merely because no one is asking for the data. In other instances, the data are not stored yet, but they can be, thanks to more easily implementable and cheaper sensors.

5.2. A Shewhartian view on safety

What does it mean to look from a Shewhartian perspective onto the field of safety? In the most generic sense it means: 1. measuring the behavior of a process that contains the potential for an incident, 2. looking at the distribution and variation of that process and 3. identifying sources of unwanted variation and eliminating them.

1. Measure the behavior of a process that contains the potential for an incident

   We have identified SPADs as an incident within the rail industry. When a train is approaching a red aspect, this is a situation with the potential of the incident, the SPAD, to occur. When a train is not approaching a red aspect, there is no potential for this specific incident to occur. As another example, in healthcare, organizations are concerned with the risk of an infection during surgery. When a surgery is being performed, this is a situation with the potential of the incident, the infection during surgery, to occur. When there is no surgery, there is no potential for this specific incident to occur.

   For every situation with the potential of a certain incident, the behavior should be measured. In the case of risk of a SPAD, we want to measure every approach to a red aspect and calculate one value per approach that summarizes the behavior. This can, for example, be the highest value of required deceleration, as will be explained in more detail below. In the case of risk of infection during surgery, we want to measure every surgery and calculate one value per surgery that summarizes that (part of the) process. The number of times the door opens during surgery is a useful measure, because the number of times the door opens correlates with the risk of an infection [99].

2. Look at the distribution and variation of that process

   When we have this one value per approach to a red aspect and we have measured 500 approaches, we can look at the distribution of these 500 values. The same

can be done for surgeries, i.e. the amount of times the door opens during one operation. The distributions will show the variation within the process.

3. Identify sources of unwanted variation and eliminate them

Figure 23a shows incidents as part of the tail of the distribution (i.e. the red area). **Figure 23b** shows the same type of distribution with the same mean, but a reduced variation, leading to a reduction in the amount of incidents. When variation is unacceptably high, as in **Figure 23a**, the sources of the variation should be identified and eliminated to reach a better situation, as in **Figure 23b**.

Finding sources of variation contrasts with the "traditional" view on investigation with a focus on the identification of causes of incidents that have occurred. Some of the factors identified in traditional investigations can be sources of variation, but the focus on variation takes a broader perspective and also includes factors that sometimes lead to outcomes better than expected as part of a large variation. The identification and elimination of sources of variation also provides additional possibilities for safety interventions other than just adding safety barriers.

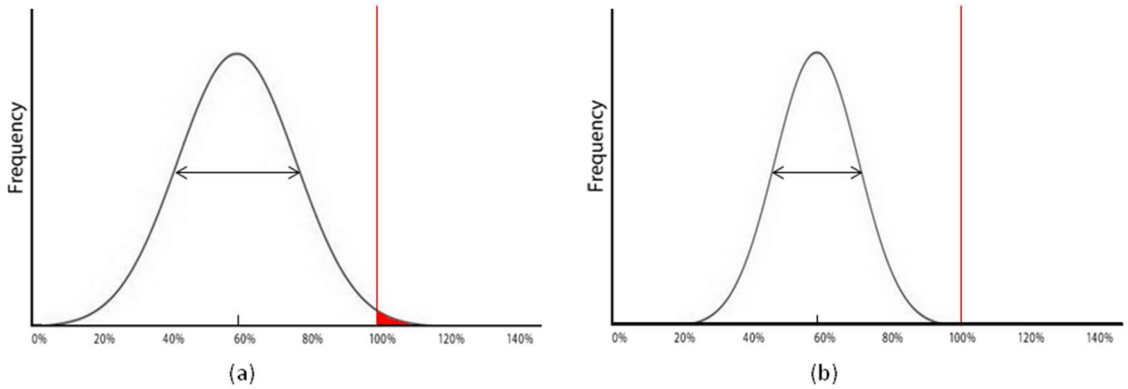


Figure 23. **(a)** original variation with incidents in the tail of the distribution **(b)** reduced variation with no incidents

Before we elaborate on approaching train driver behavior and SPAD probability via the Shewhartian perspective, we will first delve deeper into the notion of variation.

5.2.1. Two types of variation, from four perspectives

All processes contain variation. As a neutral example we will use "cooking a dish". If you cook the same recipe fifty times, it will never be exactly the same dish. If the same dish sometimes tastes awful and sometimes delicious, then there is a lot of variation. If the dish always tastes average to good, then there is a lot less variation. Even in very exact, machine driven processes, there is always variation if you measure in enough detail. A piece of metal might always be cut to 1.200 cm by a metal-cutting-machine, but when the resulting pieces of metal are measured in ten or more decimals, they will not have exactly the same lengths.

5.2.1.1. Variation aspect: problematic or not

When one cooks the same recipe fifty times, and it always tastes average to good, then one might consider this process having little variation. However, the same small differences in dish taste, might be considered a huge variation in a five-star restaurant. Whether variation is problematic or not, is therefore not only related to the absolute size of the variation, but also to the context and goals.

5.2.1.2. Variation aspect: controllable or not

One source of variation in the taste of the sauce we are making, is the amount of water that is added during cooking. One might add 120 ml one day and 140 ml the other day, because it is done by hand. One might reduce this variation by using a measuring cup. Using the measuring cup could reduce the variation in water added to 130 ml on average and ranging between 125 and 135 ml. The variation is thus controllable. The variation might be reduced even more by using a more exact measuring cup that can be read more precisely. Eventually it will no longer be practically feasible to reduce the variation (e.g. due to fluctuations in temperature affecting the density of the water being measured).

5.2.1.3. Variation aspect: common or special

Shewhart introduced two types of variation he called *chance* and *assignable*, which were later adopted as *common* and *special*. When we cook our sauce on an electric stove at level 5, then the temperature of the stove will always differ a bit, around an average. This is common variation. When the stove has an error, the temperature might be reduced to zero. This variation in temperature is not caused by common variation, but is considered special variation. Common variation is always present, whilst special variation is caused by a factor which is not always present. In the previous example, the difference in water added by hand is usually part of the common variation. When a lot more water is added one day because someone bumped into the cook, then this is special variation. One could also argue that the variation itself is not common or special, but its source is.

5.2.1.4. Variation aspect: known or unknown

Both variation and its cause can be either known or unknown.

If, whilst cooking at home, the partner of the cook adds varying amounts of pepper without the cook knowing he adds pepper at all, this will create variation in the level of spiciness of the dish. To the cook, the variation will be known when (s)he tastes it, but the cause of the variation will be unknown to him or her.

If the partner adds pepper every time the recipe is cooked, then it will be *common* variation. If the partner is only home occasionally when the recipe is being cooked, then it will be *special* variation. In both cases, the variation is known, but the source is not.

If a source of variation has not occurred yet, then both the variation and its source might be unknown.

It is also possible that a source of variation is known, but because it has not occurred frequently yet, it is unclear how this variation behaves. For example, if the partner was only present once and added three twists of the pepper mill at that time, then it is unknown whether (s)he will also add around three twists every time or whether (s)he adds six twists one time and only one at another occasion. This makes it difficult to model and predict the variation in these situations, even when the source of variation is known.

	Variation known	**Variation Unknown**
Source known	Ideal situation	Source of variation can be investigated further to understand how much and what kind of variation it causes
Source unknown	The known variation can inspire a search for the source	"We don't know what we don't know"-situation. Can be due to lack of measurement or due to infrequently or not yet occurring sources of variation

Table 9. The four types of variation one can encounter in processes.

5.2.2. Gaining more insight into your variation

It might be tempting to start the examination of variation with the question whether the variation is problematic or not. In theory, one might use the distribution characteristics to calculate what the chance is of obtaining unacceptably high values based on the observed distribution. However, this type of "black box" approach does not improve our understanding and the level of safety.

Being in full control over a process does not just mean having a process behaving within certain boundaries, but also understanding the common and special variation within that process. In the "black box" approach, minor data oversights and changes within the process can also easily lead to unexpected risks. We therefore recommend asking the question whether the variation is problematic or not after the variation and its source are better understood.

A useful question to summarize an organization's level of understanding of the variation is "Can we account for the variation?" In some cases, there will be clear signs of unaccounted variation. For example, if a process follows a bimodal distribution, where one expects a unimodal distribution, more investigation is needed and the data can be used to guide investigations in the right direction.

In other cases, the data might follow the expected distribution or there might not be a clear expectation. What should you do in these cases? In these cases, it is recommended to split the data. Splitting the data can be used to check the assumption that the variation is the same across groups/categories or to check the assumption that certain groups/categories have more variation than others. We will first apply this approach to SPAD risk and afterwards illustrate the application of this approach within health care and the fire brigade.

5.3. Trains passing red aspects

During a train drivers formal training, much time is devoted to stressing the importance of avoiding SPADs. Train drivers are fully aware of the risk of SPADs and also want to avoid these at all costs. No train driver wants to have a SPAD. Yet these incidents still occur and measures to prevent them based on "identified causes after an investigation" have worked in the past, but failed to reduce the number even further. Following Shewhart, the alternative approach would be to look at the causes of variation in train driver behavior and treat the SPADs as "extreme outcomes" of a system not optimally controlled. The first step is to determine what kind of indicator can be used to describe the behavior of the driver.

As described in section 2.2, ProRail had already developed a proactive safety measure called Time-to-SPAD (TtSPAD) in cooperation with Dutch Railways (NS) at the start of the PhD. During the analysis of the measure as part of my PhD research, we developed a new safety measure called Deceleration-to-SPAD (DtSPAD) based on the previous one.

The DtSPAD or 'required deceleration' can be calculated for any train approaching a red aspect. It indicates the deceleration that the train needs in order to still be able to stop in front of the red aspect, i.e. the minimal required deceleration that the driver needs to brake with continuously. There is a clear link between the measure and the incident (a SPAD), since a required deceleration that is higher than the total available braking power of the train means that the train will pass the signal at danger (unless the signal clears before the train reaches it).

The highest required deceleration measured during an approach can be taken to illustrate the smallest buffer the train driver has before he reaches the red aspect. This maximum required deceleration (mDtSPAD) is the one summary value per situation with a potential for the SPAD incident, (i.e. per approach to a red aspect), which we can use to inspect its distribution, and variation. For more information on mDtSPAD calculation and a graph showing the relationship between DtSPAD and actual deceleration, see section 2.2.

5.3.1. Examining current variation

The value of our train driving measure is not expected to be the same for every approach. Some variation is to be expected since train driver behavior can be affected by multiple factors, like scheduling and time pressure, energy efficient driving and personal style. We might assume that these kinds of factors are present for every red aspect approach and thus common variation. Based on knowledge of the rail system and human factors, we also expect some sources of special variation. A red aspect can, for example, be positioned immediately after a planned stop at a platform, or not. In the first situation, the train driver will stop at that location regardless whether the aspect is red or not, whilst in the latter situation the train driver will have to decelerate to standstill solely because the aspect is red. We expect different levels of variation for these two types and therefore split the overall distribution into two distributions according to that factor.

An additional factor of interest is which yellow aspects are shown before a red aspect. In Dutch rail, multiple different yellow+number aspects are used (Yellow+8, Yellow+8 Flashing, Yellow+6, Yellow+6 Flashing, etc. See section 2.1.2. for more on information on these aspects). Although there are some differences between these various yellow aspects and the accompanying infrastructure, in general the same amount of variation in behavior is expected per aspect. In other words, the signaling system is not meant to create large differences in our measure. We check this assumption by splitting the previously split data further into yellow aspect categories. The results showed that the

distribution and the variation of our train driving measure is not the same for each type of yellow aspect, which is contrary to what might be expected from the signaling design perspective and design intentions. There was one yellow+number aspect in particular, that correlated with a lot more variation.

Multiple theoretical explanations are possible as to why this particular yellow aspect generates more variation in DtSPAD. Delving further into the distribution by splitting it according to additional theoretically relevant factors (permitted track speed at location of yellow+number aspect and yellow-red distance), we discovered that this type of yellow aspect only showed (a lot) more variation in DtSPAD when certain other elements were also present.

The fact that there was a subset within the distribution of this deviating yellow aspect that showed little variation in our measure, made some of the hypotheses less likely as to why this yellow aspect might be problematic. Inspecting those situations with little variation was interesting in itself and provided insight into what train drivers are exposed to and, therefore, what they can get used to. When employee behavior is involved, learning can occur during situations which they experience often. If the situation changes slightly, different behavior might be required, but the learned behavior is nevertheless activated, as our data-splits seem to indicate.

By splitting DtSPAD according to the type of yellow variants, and adding other factors, we identified new potential SPAD-causes related to the sources of variation in the current way of driving and we could exclude others.

5.3.2. Examining change and whether it is problematic

To examine whether a process remains in control, Shewhart's control charts can be used. Alternatively, when a process innovation is tested over a period of time, one can compare the distribution of the output before and after the change to see whether it has resulted in a change in the distribution of the output.

It should be noted that not all changes are problematic. For example in the case of approaches to red aspects, higher values in required deceleration are not per definition problematic. In fact, very low values indicate that trains approach signals at much lower speeds than possible within the rail infrastructure and this potentially hampers capacity, whilst a somewhat higher speed is not necessarily less safe. When the entire distribution moves to the right (i.e. the average required deceleration increases), but the variation does not increase or only a little, it is a valid question whether this change is problematic or not.

Because an increase in our measure does not linearly relate to an increase in the probability of a SPAD, it is useful to identify a threshold for what indicates a problematically high value. To detect a change in the probability of a SPAD we can then monitor both:

1. Whether there is an increase in the amount of actual measurements with a value above the threshold.
2. Whether there is an increase in the theoretical probability of obtaining values above the threshold, based on the distribution parameters.

Option two is especially useful when there are not a lot of measurements or the probability of a high value is low, but still of interest. This approach does require knowledge of the distribution and variation in order to extrapolate.

5.4. Exploring some more applications

The principle of splitting variation into meaningful bits and deliberating whether we want that extra variation or not can also be applied to other domains than rail. Here we will present two hypothetical examples: infection during surgery and fire brigade turnout time.

5.4.1. Infection during surgery

Any surgery that causes a break in the skin can lead to an infection. These infections are called surgical site infections (SSIs). Healthcare organizations are concerned with the risk of SSI's as a result of exposure to bacterial clusters. One of the factors contributing to the growth of bacterial clusters is laminar air flow as a result of opening doors during surgery [99–101]. It is of course not possible to determine exactly which door opening caused the growth of a bacterial cluster. Neither is it possible to completely eliminate door openings. What is possible, is to determine what the variation is in the number of times a door is opened during surgery and identifying sources of the variation. The level of variation in the number of door openings is actually quite high, even for the same type of surgery, and as such forms a promising target for improvement [102].

A hospital can count the amount of times the door opens per surgery as an outcome measure. The resulting distribution of door opening frequency can be split into categories that the hospital staff expect to have the same variation or categories that they expect to have differing variation. The distribution can for example be split according to different hospitals or rooms, where one might expect the variation to be the same. In this manner, causes such as room design, poor planning or equipment failure might be identified or intrapersonal variations of behavior [103].

Mears, Blanding, and Belkoff show that the longer the surgery, the higher the number of door openings [104]. Duration of the surgery can also be taken as an outcome measure and variation in duration of the surgery can be examined. If the same type of surgery varies a lot in duration time, then identifying the sources of that variation can indirectly help in identifying sources of variation in amount of door openings.

5.4.2. Fire brigades' turnout time

When there is an emergency, the response time is the time between the moment a call comes in, to the moment the emergency services arrive on site. This response time can vary a lot and consists of multiple stages. In case of an emergency requiring the fire brigade, part of the response time consists of the turnout. The turnout time is the time between the moment the fire station receives the notification and the moment the firefighters depart from that station [105].

For each response, we can calculate the turnout time. The National Fire Protection Association, in their 2010 version of the NFPA 1710 standards, recommend a turnout time objective of 80 seconds (for fire and special operations response only) [106]. The Fire Protection Research Foundation (FPRF) reports that "The time actually required and recorded for turnout of 90% of the calls was 123 seconds for fire" ([107] , p.17).

A fire station commander might consider 170 seconds too long and measure turnout times for two months. After two months, no values over 170 seconds have been measured. Given the distribution of the turnout times, we can calculate what the theoretical probability is of a value above 170 seconds as a result of the variation. This

calculation can indicate a too high or sufficiently low probability, but it is more interesting to consider the variation of the turnout times.

Reglen and Scheller state that the time of day has a large impact on the turnout time, with turnout times being longer during the so-called graveyard shift [108]. The graveyard shift is defined in their study as midnight to 6 a.m. Splitting the distribution into daytime shift and graveyard shift could reveal whether this effect is also present in the station of our commander and whether only the average is higher during the night with the same variation or with a difference in variation.

To examine the variation in more detail, the fire station commander might expect that the variation in turnout time is the same for different team compositions. The distribution can be split according to team compositions and the assumption checked whether the amount of variation is indeed the same. Differences in variation should, however, not be used as direct evidence that certain employees, teams, or situations are "worse" than others. The distributions are split to account for the variation and look for gaps in the knowledge about turnout practices. Differences between two teams can appear to be present when they differ in other aspects as well; for example, when one team has more graveyard shifts. The variation caused by time of day should be taken into account, as well as the possibility for other differences.

5.5. Checking your variation: some afterthoughts

5.5.1. Similarities to other approaches

The approach of splitting is very similar to the testing of hypotheses in statistical procedures like ANOVA. Differences are that there is a focus on *variation* and not *means*, which is often the case. Especially in comparison with more advanced prediction models, there is a focus on understanding the data, so that it can inspire identifying new potential incident causes. It will also lead to more informed decision making. For example, if a specific subset correlates with higher risk values, then it will be unwise to increase the occurrence of this subset, without first addressing the cause, whilst increased occurrence of a different subset with lower risk scores might be a better choice to attain certain business goals without compromising safety. "Split", or subset distributions can also be used for better early warning indications, as used in the control charts by Shewhart, if they differ from the overall distribution.

5.5.2. Criticism on Shewhart's statistical process control applied outside of quality management

5.5.2.1. Criticism: Not all variation is controllable

Shewhart advocates a reduction in variation, but variation might not always seem controllable, especially when it comes to situations with human behavior. Whilst it is true that not all variation is controllable, this is not a reason to ignore the variation that is controllable or possible to influence. Much human behavior is, for example, not fully controllable, but (physical) circumstances have been shown to have a large impact on human behavior [30]. Influencing these circumstances can lead to a reduction in variation in human behavior.

Even if a source of variation turns out to be beyond control, knowledge of this uncontrollable variation can still be useful, because if we can account for that variation, it can reveal remaining unaccounted variation that is controllable. Additionally, even if a factor causing variation is uncontrollable, implementing the process is still a choice and

one might choose to downscale or not upscale if that factor is prevalent. Knowledge of the uncontrollable variation can also be a reason for implementing additional protective measures.

5.5.2.2. Criticism: It is not just about variation, but also about the mean

Within the Shewhart approach, the focus lies on influencing the variation with the mean taken, either implicitly or explicitly, as a given. From the perspective that incidents are extreme values of the normal distribution, it is a valid point that the mean should not be forgotten. In fact, all parameters that make up the distribution (a distribution could also be binominal), are relevant. The Shewhart approach advocated here does not exclude influencing the mean, but proposes examination of the variation and distribution first, whilst more traditional approaches tend to put the mean first and forget or underemphasize the other distribution parameters. After all, when there is a lot of variation, the mean gives us very little information, yet when the variation is low, the mean gives us a lot of information.

Examining variation can also be a step towards influencing the mean. Once sources of variation are identified, it will also become clearer which factors or circumstances lead to better outcomes. For example, team size can be a source of variation in our outcome measure with small teams having better performance outcomes than large teams. The organization can use this knowledge to change the mean of its performance, in this case by using smaller teams more often.

5.5.2.3. Criticism: It should not be about variance, but about incident potential

This chapter focuses on the probability of an incident occurring, when there is incident potential. One might argue that it is a more thorough approach to eliminate the incident potential. For example, in the case of the risk of a SPAD, one might wish to reduce the number of red aspect approaches instead of reducing the risk of a train driver passing a red aspect when it is present.

Reducing incident potential is not always possible, though. To prevent a leak, an organization might theoretically stop storing products, but this might put them out of business. An additional aspect to consider is that reducing the number of situations with incident potential can affect the remaining situations with incident potential, as employees are no longer trained on the job on how to deal with these situations. This can be compensated by creating similar situations without the risk, like fire drills or training in simulators, but this might not be feasible or effective. It is therefore worthwhile to examine both possibilities in the organization's attempts to reduce incidents.

5.5.3 Taking the Shewhartian approach another step further

5.5.3.1. From "right" and "wrong" to "deviations"

The main focus of this chapter is on measuring processes and examining their variation. The core notions of Shewhart's take on variation can also be taken a step further, outside the realm of numbers. Taking the Shewhartian approach means that the focus will shift to any kind of deviations from the intended process, either "too bad" or "too good". In essence, there is a shift in focus from the question "what is right and what is wrong" to "why are there differences? Why are there deviations from the expected?"

In this perspective, even positive outcomes can be a reason for investigation, because the very existence of the variation is a cause of concern, although the outcome might occasionally be positive. If a job takes on average 30 minutes but for some worker it takes an hour, this would probably raise concerns about the competency of the worker and the quality of the work. If the same job is done in only five minutes the reaction is

likely to be "that's fantastic!", while it should be: "that's worrying". Any deviation is considered worth reporting and investigating because it shows a lack of control and that is a cause of concern.

Rather than stimulating employees to report only on "what is wrong?" the scope would be expanded to "what is different from the expected?". Suppose a nurse detects that a piece of equipment is missing. In the "traditional" approach, it would first be considered whether the absence of this particular tool is safety critical and whether the increase in risk warrants action. Only after the answer is affirmative, the absence is reported. According to Shewhart's approach, the absence is a deviation from the intended process and will create avoidable variation, and that is a reason to take action. Sammer, Lykens, Singh, Mains and Lackan state in their review study on patient safety culture that in a "good" safety culture: "standardization to reduce variation occurs at every opportunity" ([109], p.157).

5.5.3.2. Boring jobs?

It may seem that the aim to reduce variance and deviations makes working life boring and reduces the need for craftmanship, but that is not necessarily true. Being able to detect deviations is a task that requires craftsmanship and which, if the reporting is positively reinforced, can create extra job satisfaction. In most organizations, there is also enough "uncontrollable variation" that requires craftmanship to deal with. For example, there is a bewildering range of characteristics of patients in hospitals that cannot be controlled: young, old, obese, not obese, male, female etc. They all have slightly different symptoms and a wide range of expertise is necessary to treat them all. What a surgeon (and the patient) does not need is extra variation as a result of poor communication, absent or broken equipment and organizational disorder. So, although medical professionals have been very good at compensating for these kinds of deficiencies and save patients from serious harm, this is an unsustainable situation in the long term. Organizations should not need "heroes" to save the day; they should provide their employees with predictable processes that allow them to make optimal use of their craftmanship.

5.6. Conclusion

One of the big challenges organizations face with respect to process safety, or any other type of safety, is answering the question: Are we in control? The presence of incidents indicates that the organization might be partly out of control. The absence of incidents, however, does not equate to being in control. We advocate a stronger emphasis on the identification and elimination of pertinent process variations as the next step in improving operational and safety performance.

Some organizations might feel overwhelmed and lack the infrastructure to process this kind of information and take the necessary actions. The short-term investments should not be underestimated but in the long term the effects of such an approach are substantial and in a high-risk industry like hospitals and railways, unavoidable. Both in trying to understand and account for the variation, followed by potentially reducing the variation and taking it into account during monitoring for early warnings. The technology is ready for these kinds of implementations for (process) safety purposes. The next step is to create organizational support to take this approach on board to help reduce the number of incidents even further.

Chapter 6

HF perspective on big data tasks: Identifying cognitive pitfalls in the verification step

Based on the article: Burggraaf J, Groeneweg J, Sillem S, van Gelder P. How Cognitive Biases Influence the Data Verification of Safety Indicators: A Case Study in Rail. Safety. 2019; 5(4):69. https://doi.org/10.3390/safety5040069

Chapter summary

Using big data effectively is not a straightforward process. There are many decisions that have to be made about topics including data collection, data (sub)selection, data verification, data analysis methods and parameters and output visualization and presentation to support the data interpretation. Human Factors specialists using big data to investigate human factors topics should be aware that they themselves are also susceptible to making errors. This raises the question: can 'we' (human factors experts) also use our knowledge of human factors to decrease the probability of errors during the use of big data?

In this chapter we look at the task of data verification. Depending on the initial data quality and the approaches available, data verification can vary in complexity. During my own research, the data verification process consisted of checking and judging the data. This was a cognitively complex task. Within cognitively complex tasks, cognitive biases are likely to occur and can lead to errors. These errors can cause analysists to overestimate the quality of the data and safety experts to base their decisions on data of insufficient quality.

Cognitive biases describe generic error tendencies that arise because people tend to automatically rely on their fast information processing and decision making, rather than their slow, more effortful system. This chapter describes five biases that were identified in the verification of our behavioral indicator mDtSPAD. The insights and recommendations in this chapter can help improve data verification processes.

The additional message is that there is also value in examining the use of big data from a Human Factors perspective. In this chapter we only look at the data verification but other tasks such as the visualization and interpretation of big data results could also benefit from examining them from a Human Factors perspective and improving the employee-task interaction for researchers and analysists as well.

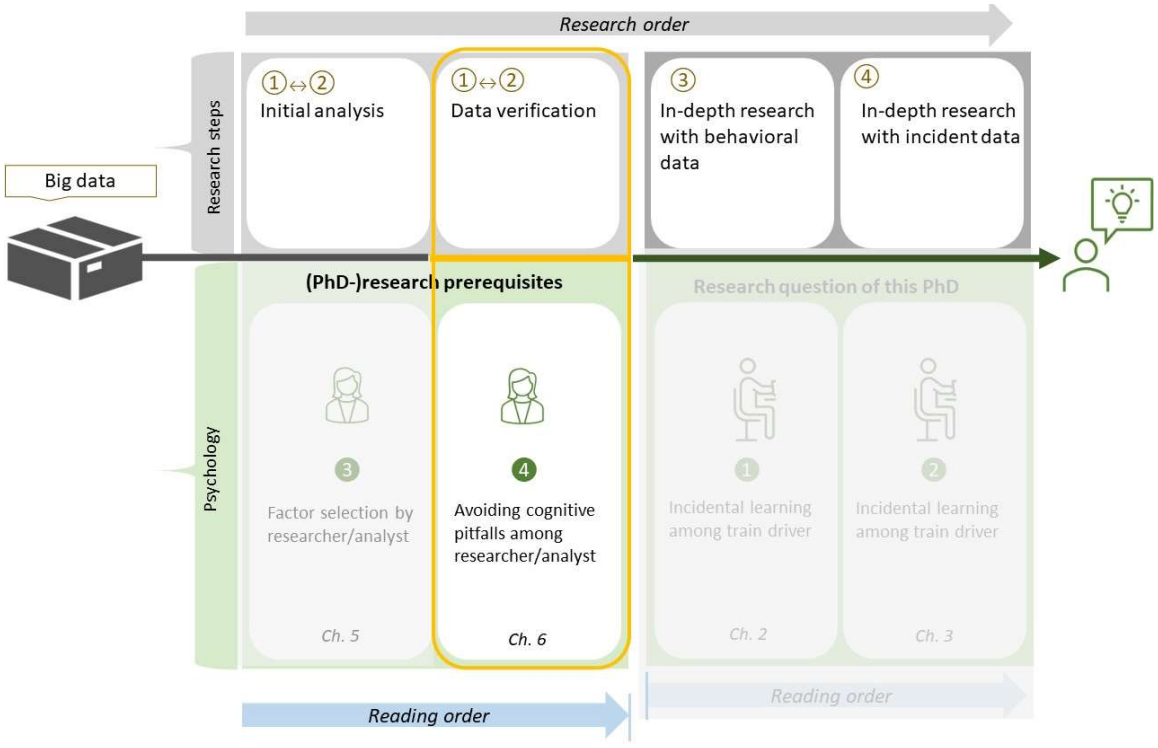


Figure 24. Location of chapter 6 with respect to research order and reading order. Chapters 2 to 4 discussed the research question of this PhD. Chapters 5 and 6 discuss the research prerequisites.

6.1. Introduction

The field of safety and incident prevention is becoming more and more data based. Organizations and institutions gather and analyze more data than ever before. Representatives from many different professional domains seek the benefits of the technological developments. Most are already implementing (big) data methods ranging from the traditional statistical analysis to state-of-the-art artificial intelligence and deep learning. Within the field of safety, new safety indicators can be used to find more detailed incident causes and effective solutions.

The field of safety however tends to have a constraint that is not shared by all fields: The data quality needs to be high. Decisions that are made can literally mean the difference between life and death. When the stakes are high, certainty is a well sought-after commodity, sometimes leading to overconservative choices. Data can help support decision making to create a better bridge between safety and innovation. This can be done by finding the common ground of overall improved execution of the core business, but only if the data can be, is and should be trusted.

Many examples unfortunately show that good data quality is not a given. Problems of faulty input data or algorithms can go undetected even when they occur frequently, like the following two bugs in software programs: "A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions" [110] and "we found that the most common software packages for fMRI analysis (SPM, FSL, AFNI) can result in false-positive rates of up to 70%" [111].

There are multiple estimates available of the number of software bugs per number of lines of codes. Whilst the exact ratio estimates vary, it is generally accepted that there is such a ratio and when the number of lines of code increases, so do the number of bugs [112]. Some of these bugs might not affect the outcome significantly, while other software bugs can have large consequences, like the infamous and expensive bug in the software of the Ariana 5 rocket leading to a disintegration of the rocket 40 s after launch [113].

Besides software problems, unreliable information or input can also lead to the publication of incorrect results. Medical investigators have later learned that the cells they studied were from a different organ than expected. Such basic specification problems are not solved by having larger sample sizes [114]. Out-of-date documents can be a cause of errors, for example if a stop sign is moved but the documents are not updated to include the new location. Errors can also occur at a later stage, for example during data integration. Data integration has become more difficult due to a larger range of different data sources containing different data types and complex data structures [115].

The impact of low data quality can be very high, depending on the use case. If emergency services are sent to an incorrect location, the consequences can be negative and immediate. If incorrect data is used as a basis for performance indicators, the effects might not be immediately visible but can still be negative. When the indicators are used for safety related decision making, unsafe situations might appear safer than they are. On the other hand, safe situations can appear problematic, leading to unnecessary or even counterproductive measures. Especially in the era of Big Data, there is increased potential to draw erroneous conclusions based on little other than volume of data [116].

6.1.1. Verification Is Complicated

The above examples show the value of successful verification. Checking and judging data is however a complex task. Quality is a multi-dimensional concept including, amongst others, accuracy (both noise and errors), consistency and completeness [115]. At the moment, more data of varying quality is being collected and used than before [117]. Additionally, data consumers used to be (directly or indirectly) the data producers in many cases, whilst currently the data consumers are not necessarily also the data producers. This is because of the large range in different data sources that are used.

The large data volume is also a challenge, both in amount of data per variable and in the high number of variables that can be integrated. When variables are being computed based on multiple sources, then verifying the quality of the individual sources is not sufficient. Verifying the computed variables alone is also not sufficient as problems can become less visible after sources are combined. Overall, verification can be a complex task for many reasons. Section 6.2.2 gives an overview of verification activities performed for the DtSPAD measure.

6.1.2. Cognitive Biases as Problem

In this chapter it is hypothesized that successful verification of data is hampered by the occurrence of cognitive biases. Cognitive biases are systematic errors in judgment [118]. This type of bias causes people to err in the same direction in the same information judgment task. The existence of cognitive biases during complex judgment tasks has been confirmed multiple times within numerous different experiments [119]. Cognitive biases have also been identified specifically within the domain of risk management, namely in incident investigation reports [86] and during process hazard analysis studies [120].

A lot of research has been done into cognitive biases since the pioneering work by Kahneman and Tversky in the early 1790's [121–123]. Early research often consisted of experiments in which college students were presented with contrived questions they had to answer. As a result, it has been hypothesized that cognitive biases are an experimental artefact [124]. Research has however continued in more realistic settings and within a vast amount of topics (e.g., [125]). There is for example research on cognitive biases in specific health-compromised groups (e.g., persons with depression), different types of decision making (e.g., medical diagnosing), the negotiation process, project management and the military.

6.1.2.1. Preventing Cognitive Biases

Research on cognitive biases in specific domains can be very useful, because it is difficult to apply generic knowledge about cognitive biases to prevent errors. There are several reasons why this is difficult. First, it is not efficient to try to eradicate all cognitive biases in human cognition. The "slow" information processing which counteracts cognitive biases can come at a substantial cost. Our brains for instance consume 20% of our oxygen at rest and even greater proportions of our glucose, despite taking up only 2% of our body weight [126,127]. Trimmer [122] hypothesizes from an evolutionary perspective that cognitive biases arose for two reasons: (1) To reach optimal decision making in favor of evolution, and; (2) to reach a balance between decision quality and internal cost. Kahneman's explanation [119] of cognitive biases in terms of two systems for cognition highlights the subjective experience of effortlessness belonging to the system responsible for cognitive biases. The subjective experience of the other system is one of significant effort.

Secondly, cognitive biases cannot be prevented by simply telling people about their existence. People tend to think they are less susceptible to biases than other people, which is called the bias blind spot. Pronin and colleagues [128] found that the bias blind spot was still present even after the participants read a description of how they themselves could have been affected by a specific bias. This bias blind spot is specifically related to recognizing our own biases, while people tend to recognize and even overestimate the influence of bias in other people's judgment [129]. Whilst extensive training in recognizing one's own cognitive biases is possible, the effectiveness is unclear and the training could be very expensive.

Another option is to redesign the person-task system to inhibit the bias that interferes with the task (Fischoff, 1982 as in [123]). Planning poker is for example an estimation technique which has been specifically designed to prevent anchoring bias. Participants independently estimate for example "required time" or "cost" for a task and then simultaneously reveal their estimates. In this way, there is no anchor to be influenced by as there would have been if a number was spoken out loud by one person before others had made their estimates [130].

In the previous example, the problem of incorrect estimations in project planning was traced to being (in part) caused by a cognitive bias and debiasing action was undertaken. It is of course not always known which problematic errors are present within an organization or department. Errors might not be reported or recorded and especially in the case of errors as a result of cognitive biases, they might not even be noticed. Cognitive bias theory can be used to predict which errors might occur in specific tasks and thus help identify errors that are likely to reoccur. Both knowledge of cognitive biases and the specific tasks can then be used to redesign the person-task system.

Research on cognitive biases in specific domains can thus be very useful. A search in the web of science database yielded few articles about both cognitive biases (or human factors) and big data. On the other hand, there has been some research on cognitive biases in software engineering. While this field is obviously not the same as big data, it does contain some tasks with parallels to the verification process, specifically the testing of the code. The review by Mohanani and colleagues [123] provides interesting insights: The earliest paper of cognitive biases in software engineering was published in 1990, followed by one or two papers per year until an increase in publications as of 2001. Mohanani and colleagues found that most studies employed laboratory experiments, and concluded that qualitative research approaches like case studies were underrepresented. Most studies focus on the knowledge area software engineering management, whilst many critical knowledge areas including requirements, design, testing and quality are underrepresented.

The next sections of this chapter describe the method we used in our study and the identified biases. The remainder of this introduction will first be used to explain what cognitive biases are and what the generic mechanism is behind this specific type of errors. Knowledge of this mechanism helps to understand the chosen methodology and the five cognitive biases that will be discussed in the results section.

6.1.2.2. Cognitive Biases: System 1 and System 2

Burggraaf and Groeneweg [86] (pp. 3–4) clarify the mechanism behind cognitive biases as follows: "According to the dual-system view on human cognition, everyone has a system 1 (fast system) and a system 2 (slow system), also known as the hare and turtle systems. Our system 1 generates impressions and intuitive judgments via automatic processes while our system 2 uses controlled processes with effortful thought [118]. System 1 is generally operating, helping you get around and about quickly and without

effortful thought. Questions like "1 + 1 = …?" or "The color of grass is...?" can be answered without a lot of effort. The answers seem to pop up. When our system 1 does not know the answer, our system 2 can kick in [119]. System 2 requires time and energy, but can be used to answer questions like "389 times 356 = …?" The switch between system 1 and system 2 based on necessity, is an efficient approach. The problem is however that system 1 often provides an answer, even though the situation is actually too complex. We often think the answer from system 1 is correct, because it is difficult to recognize the need for system 2 thinking when system 1 answers effortlessly, but this is actually when a cognitive bias can occur. The main problem leading to cognitive biases is therefore not that people cannot think of the right solution or judgment (with system 2) but that people do not recognize the need to think effortful about the right solution. This lack of recognition also explains why making cognitive biases is unrelated to intellectual ability [131].

6.1.2.3. System 1: Automatic Activation

One of the mechanisms underpinning system 1 is the automatic spreading of activation that occurs within the neural networks of our brain. The spreading activation theory postulates that whenever a concept is activated, for example after seeing it or talking about it, this activation automatically spreads out towards the other information that the particular concept is related to [132]. This automatic spreading of activation can lead to cognitive biases when irrelevant information is activated and/or not enough relevant information is activated [118]. This follows the description of judgement biases "as an overweighting of some aspects of the information and underweighting or neglect of others" ([118], p. 1).

Information or knowledge is not stored randomly in the brain but in meaningful networks, with related concepts close to each other. The information that is more closely related to the concept becomes activated more strongly than the information that is less closely related to the concept. When information is activated in the brain, the chance of thinking about it is increased [132]. We can for example activate the concept of the animal sheep in your brain by talking about sheep and how they walk around, eat grass and bleat. If we would now ask you: "Name materials from which clothing can be made," we can predict that you will think of wool first, before thinking of other materials, because it was already slightly activated along the concept of sheep. Some other materials might come into your mind via system 1 quite quickly as well, while you will have to search effortful with system 2 to think of final additional options.

The mechanism of automatic activation in the context of cognitive biases is clarified further below by taking the confirmation bias as an example. The confirmation bias describes the process in which people search for, solicit, interpret and remember information that confirms their hypotheses and discount or ignore information that disconfirms them.

The confirmation bias is caused by information processes that take place more or less unintentionally, rather than by deceptive strategies [133,134]. When testing a hypothesis, the activation of the hypothesis increases the accessibility of information in memory that is consistent with the hypothesis [135].

For example, when one considers the folk wisdom that opposites attract, multiple examples of couples of two different people are automatically activated, and the person judges the folk wisdom as true indeed. Or multiple examples are activated of how you and your partner are different and yet so good together. However, when one considers the folk wisdom that birds of a feather flock together, multiple examples of couples of two similar people (perhaps even the same couples as before, but now with respect to

different parts of their personality) are automatically activated, and the person judges the folk wisdom as true indeed. Counterevidence for each piece of folk wisdom is not automatically activated, because it is not close to the activated concept in the network of activation. To activate counterevidence, one must actively think of counterevidence and thus use his or her system 2.

The enhanced activation of confirming information also influences the perception of other confirming information, which is then easier to process and activate. One can for example read an article with two consistent pieces of information and two inconsistent pieces of information and yet feel that the author's hypothesis is supported as the consistent information is processed and remembered more easily, without the need for effortful thought [136].

A countermeasure is to think of alternative scenarios, alternative hypotheses and a good old-fashioned dose of effortful thought. Multiple experiments on biases have shown that the instruction to retrieve incompatible evidence did indeed alter judgment, while instruction to provide supporting evidence which was already automatically activated, did not alter judgment [118].

6.1.2.4. Relation between System 1 and System 2

For explanation purposes, the terms "system 1" and "system 2" were used. It is important to note, that in this dissertation, they are not considered as separate independently operating systems. The automatic spreading of activation as part of system 1 is a core functioning of the brain and shall always occur. It might not always be sufficient to lead to a direct answer, but the mechanism is present. Preventing cognitive biases is therefore not a matter of trying to switch off system 1 thinking, but of adding system 2 thinking, which means activating other relevant knowledge apart from the automatically activated concepts.

It is not possible to suppress the automatically generated activation. The two images below are meant to illustrate this. Both images (see **Figure 25**) contain the capital letter A. When seeing only **Figure 25a**, it tends to be hard to see this letter. More noticeable are other patterns like the clustering of yellow on the bottom left and the wrinkly line through the middle. In **Figure 25b**, containing the exact same ordering of the circles, it is very easy to see the letter A. When people know the "correct answer" after viewing **Figure 25b**, they are able to see the letter A in 19a, but still find it quite hard to suppress the other patterns. These other patterns tend to "compete" while one tries to see the letter A. It is very hard to ignore the irrelevant information, even when you know it is irrelevant.

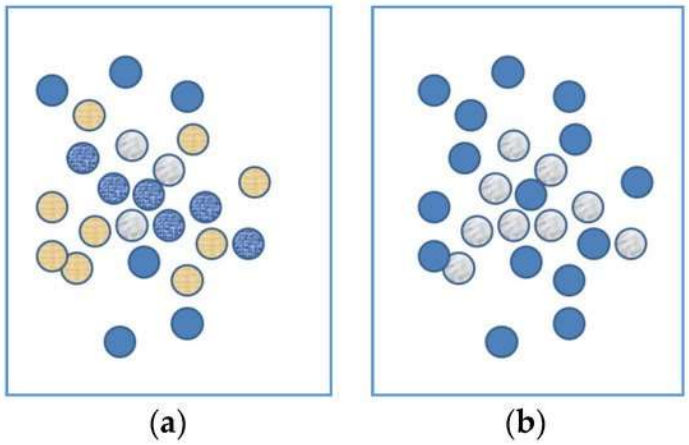


Figure 25. (a) Image containing capital letter A; **(b)** Image containing capital letter A.

This is one of the key elements in identifying cognitive biases. The first cue is that they are errors which we tend to make repeatedly. When we are capable of giving the correct answer, it is not because we are able to prevent the thought that feels intuitive, from occurring, but are able to correct it with a more reasoned thought. It is this pattern of being tempted to give an incorrect answer, which is characteristic to these types of errors.

So far, we have talked about generic cognitive biases which occur across domains. The mechanism behind these biases can also lead to more specific errors. These more specific errors are the result of the general mechanism in combination with specific knowledge about a domain, or domain-specific associations. The domain-specific manifestations of the biases will be called "cognitive pitfalls" from this point on. The hypothesis is that there are cognitive pitfalls present in the verification process. The accompanying question is: which cognitive pitfalls can occur during the verification of data (for a quantitative safety indicator)?

6.2. Materials and Methods

Mohanani and colleagues [123] state in their review on cognitive biases in software engineering that qualitative research approaches like case studies are underrepresented, with most empirical studies taking place in laboratory settings. For the current research, the case study method was used. Yin [137] wrote in his book on case study research: "In general, case studies are the preferred strategy when "how" or "why" questions are being posed, when the investigator has little control over events, and when the focus is on a contemporary phenomenon within some real-life context." (p. 1) He goes on to say that "the case study allows an investigation to retain the holistic and meaningful characteristics of real-life events—such as individual life cycles, organizational and managerial processes, neighborhood change, international relations, and the maturation of industries." (p. 3)

The case study method makes it possible to cover the contextual conditions, which are essential for the current study [137]. One of the seminal ideas that emerged from case studies includes the theory of groupthink from Janis' case on high-level decision making [138].

In the current study, participation-observation and informal interviews were used to collect information and identify errors during the verification of a safety indicator. The identification of cognitive pitfalls was guided by theoretical propositions, specifically several criteria.

6.2.1. Method of Pitfall Identification

The method of identifying cognitive pitfalls consisted of (1) identifying errors, (2) checking whether the errors were possibly caused by system 1 thinking and (3) identifying the common ground between errors independent of the specific context, but within the verification process and (4) explaining the error in terms of system 1 automatic activation.

1. Identifying errors

The word "error" here refers to having held an incorrect belief. In order for an error to be recognized, one must realize and believe that his previous statement was not true. In

other words, an error has occurred when a person retracts their statement saying they no longer believe it is true.

2. Check whether the errors were possibly caused by system 1 thinking without system 2 compensation

Three cues were used to check whether the error could have been caused by system 1 thinking. A or B should occur and C.

A. Tendency to have the exact same incorrect belief again by the same person, despite having been aware of its incorrectness.

Cue A corresponds to the hard-wired nature of system 1 thinking and reduces the chance of the specific error manifestation being the result of randomness. For example, when there is a different error inducing factor, like time pressure, this can cause errors in a wide range of tasks and the resulting error, error A, could just as easily have occurred as error B. When error A only occurs once, this is not necessarily a reoccurring error that we as humans are vulnerable to due system 1 thinking.

B. Other people have the same incorrect belief (or had it cross their mind before correcting themselves).

This cue corresponds to the characteristic of cognitive biases being person independent, and, like cue A, reduces the chance of the specific error manifestation being the result of randomness.

C. The person had/could have had access to the correct information via system 2 thinking.

A false belief is not caused by system 1 thinking if the person simply did not have access to the correct information. For example, if a person was told that it takes three hours for a certain type of tank to fill up and he or she believes this until finding out it actually takes four hours, this person had an incorrect belief, but not because of system 1 thinking/a cognitive pitfall.

However, consider the following scenario: there are two trains approaching a signal showing a red aspect, and both trains have the same required deceleration to still be able to stop in front of the red aspect, but train A is closer to the red aspect than train B. Given that all other factors are equal, which train is at the highest risk? In this scenario someone might now answer "train A, because it is closer", but after discussion say: "In my first answer I did not consider that train B must have a higher speed than train A, therefore I don't think it is train A anymore, but train B". The rejected belief in this example can be the result of system 1 thinking, because the person did not hear any new information, only used already known information in answering the question, which he or she had not done before.

False assumptions are also a candidate for system 1 thinking. For example, one might assume that a sensor is gathering the correct data. If it later turns out that the gathered data was incorrect, then the previous incorrect assumption could have been a system 1 error. The argument "but we did not know the sensor was faulty", does not change the fact that the persons in theory did have access to the correct information. By thinking about the quality of the sensor, they could have realized that the quality was in fact unknown and could be bad. This is in contrast to for example being asked what the capital is of a country. If you have never heard or read what the capital of the country is, no amount of thinking will lead to the correct information.

3. Identifying cognitive pitfalls

When the same type of error manifestation occurs within different topics, for example with respect to different data sources, then the common cognitive pitfall is identified.

4. Explaining pitfall in terms of system 1 automatic activation

As a final step, it should be possible to explain the occurrence of the pitfall in terms of system 1 automatic activation. The explanations listed among the results in section 6.3 sometimes include schematic representations of knowledge structures and the automatic activation. These visualizations are not empirically proven within this study but included to illustrate how the theory of automatic activation can explain the occurrence of the cognitive biases. Even though it is not yet clear how exactly information is stored in our brains, being able to explain errors in terms of system 1 thinking and the automatic spreading of information is an indication that interventions tailored specifically to cognitive biases could have more effect than other error prevention approaches.

6.2.2. Verification of the Deceleration-to-SPAD

As discussed in section 5.3, the DtSPAD measure was developed to investigate SPAD probability. A DtSPAD that is higher than the total available braking power of the train means that the train will pass the signal at danger unless the signal clears before the train reaches it. Besides DtSPADs higher than 100% of available braking power, high DtSPADs can also be interesting for safety monitoring as they indicate small buffers. The maximum DtSPAD can be taken to illustrate the smallest buffer the train driver had per red aspect approach. The distribution of maximum DtSPADs can then be used to monitor train driver behavior and effects of interventions on behavior.

Variables that were used to calculate DtSPAD include:

- distance from GPS sensor to head of the train (inferred via the driving direction of the cabin with the sensor and train-type dependent possible sensor positions);
- location of the signal in longitude and latitude;
- signal aspect at given times;
- longitude and latitude of GPS sensor;
- speed of the train;
- for combining data: Train number, train type and time;
- originally needed for time calibration because of non-synchronous clocks: Time the train passed a signal according to hardware in the tracks and according to GPS sensor.

The data was gathered from existing systems from ProRail and NS, pertaining to the whole of the Netherlands. None of these systems were specifically designed or chosen with the goal in mind of calculating the DtSPAD indicator. The GPS sensors that were initially used were installed by the organization performing the maintenance of the trains with the aim to find the location of the trains due for maintenance. There were other sources monitoring train location at the time (2014-2016), but the data from these sensors was chosen because of the higher logging frequency compared to other systems providing data at the time.

One of the use cases for the DtSPAD was to identify factors that correlate with higher or lower DtSPAD values. Potentially correlating factors were therefore also verified in addition to the DtSPAD variable and the variables used for its computation.

Both qualitative and quantitative verification methods were used as recommended by Cai and Zhu [115]. The examination of the variables was done in the following manner:

- Where possible, quantitative variables were compared to a reference value. For example, the distance traveled between two points according to the GPS data was compared with the distance traveled according to the time between the two points and the speed.
- Variables were also checked for impossible or improbable values (e.g., higher speed or deceleration than the trains are capable of) or impossible combinations (e.g., low risk value, but negative distance to red aspect). Identified problematic values were not simply removed. The individual cases were examined in a qualitative manner to determine the cause and to fix the cause.
- Patterns were also examined for oddities (e.g., when 99% of the values follow a curve and some do not) and the deviating red aspect approaches investigated.

We analyzed the data in the programming language and software environment "R", using our own code. The data that was used for verification covered periods of one month up to a year. The exact period varied due to the iterative nature of the verification process in which improvements to the data source or code could sometimes not be implemented retrospectively. As a result, data from the last update up to the day of analysis was used.

Apart from examination of the variables, qualitative verification of the code itself was also performed occasionally, as will become evident in Section 6.5.

The cognitive pitfalls framework was applied to the verification process from the start of the verification in March 2016 until October 2016.

6.3. Results

Five cognitive pitfalls were identified during the verification process: "the good form as evidence"-error, the "improved-thus-correct" fallacy, "situation-dependent-identity-oversight", "impact underestimation" and "beaten path disadvantage". These pitfalls will be clarified by an example, explanation of the pitfall and examples from the case study, after which the implication of the pitfall is discussed. It is noted that this list of five is not necessarily exhaustive. It is possible that there are other cognitive pitfalls relevant for a given verification process that are not in this list because they did not occur during this specific case study or did not lead to salient errors.

6.3.1. Pitfall 1

6.3.1.1. Example

In this example we are looking at a variable which we expect, based on theory, to follow a normal distribution. We check the actual distribution of the real data as a means to check the quality of the data. The image below, **Figure 26**, is the result. What conclusion do we draw with respect to the quality of the data?

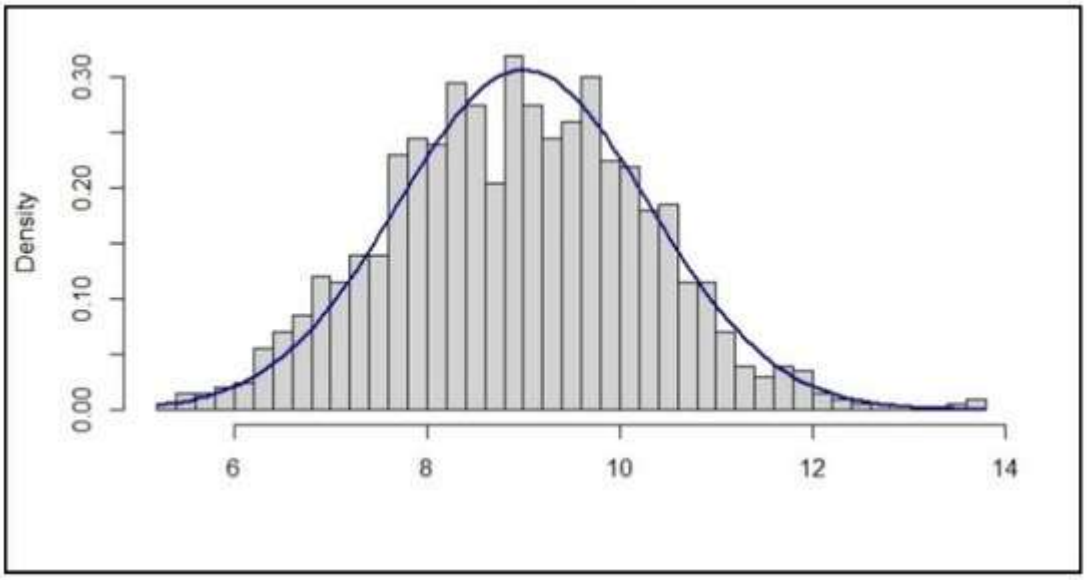


Figure 26. Example of data distribution.

A typical response would be that the data is approximately normally distributed. The data looks "about right; quite good", etcetera. Generally, this is seen as a reassurance that the data is correct and we can proceed.

6.3.1.2. The Good Form as Evidence-Error

The images in **Figure 27** roughly show all three situations which can occur when visualizing the data: (A) the data follows the distribution perfectly, (B) the data distribution looks about right and (C) the data looks awful in the sense that it does not meet expectations at all.

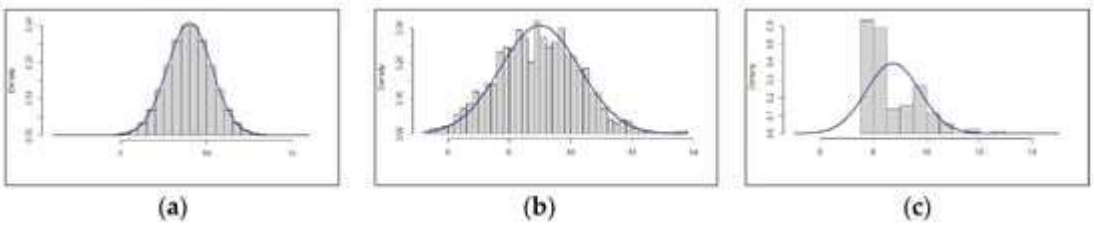


Figure 27. Types of data distributions **(a)** perfect; **(b)** good form; **(c)** ugly.

In the first case, the conclusion typically is: "That is too good to be true. This is proof that something is not right. We need to check this." As mentioned before, the conclusion in the second case typically is: "Approximately fits the distribution. This is proof that the data is correct." As well, in the third case: "This does not look at all like expected. This is proof that something is probably not right. We need to check this."

Whilst the first and third conclusion are correct and lead to the desired behavior of further verification, the second conclusion is not correct. This type of visual representation is not proof that the data is correct, since this distribution can occur as the result of correct, but also as the result of incorrect data. We tend to underestimate the chance that the underlying data is incorrect when we see this kind of "good form" visualization. Incorrect data here refers to either faulty data sources or erroneous algorithms. The actual probability of the data being incorrect when we consider the evidence of "good form" can be calculated via Bayes' theorem:

$$P(incorrect\ data\ |good\ form) = \frac{P(good\ form|incorrect\ data)P(incorrect\ data)}{P(good\ form)}$$

The actual probability includes (1) base rates of incorrect data and of good form and (2) the estimated probability of incorrect data leading to good form. In probability estimates like these, people tend to rely on representativeness and not include the base rate. This fallacy is called the base rate neglect, previously described by Tversky and Kahneman [121] as "insensitivity to prior probability of outcomes". This is possibly part of why we underestimate the presence of incorrect data in the face of "good form".

Another part of the reason can be our association between appearance and quality. We have a deep-rooted association between "bad" and "ugly" or "too perfect". Villains tend to be depicted as physically ugly persona's or too perfect persona's, usually con artists.

The strength of this association is underscored by the surprise we feel when confronted with something that does not fit this association. During the verification of the DtSPAD, we looked at a distribution of the DtSPAD variable which resembled the "good form" as previously displayed. Even after knowing that the displayed data was incorrect (because an error in the code was identified), we were still inclined to draw conclusions based on the data we saw. The notion that bad data could look like good data remained counterintuitive, while the intuitive association automatically gets activated: "but it is good looking data, so good quality data."

In reality, it is possible that bad data looks good. Even though we do not know the numbers to the base rates or relations, we can enter hypothetical data in Bayes' theorem to get a feel for the actual probability of incorrect data when visual inspection shows "good form".

$$P(incorrect\ data\ |good\ form) = \frac{P(good\ form\ |incorrect\ data)\ P(inccorect\ data)}{P(good\ form\ |incorrect\ data)\ P(incorrect\ data) + P(good\ form|correct\ data)P(correct\ data)}$$

In the first draft of an indicator, let's assume that the base rate of incorrect data is high, say 0.7. When incorrect data leads to good form with a chance of 0.3 and correct data leads to good form with a chance of 0.95, the probability is:

$=0.3 \times 0.7/(0.3 \times 0.7 + 0.95 \times (1 - 0.7)) = 0.21/0.495 = 0.42$

This example indicates that it is actually highly likely that data is incorrect, even though it looks good.

Even when we assume that incorrect data only leads to good form in 10% of the cases, the probability of the data being incorrect in the face of "good form" is still relatively high (0.20).

6.3.1.3. Implication

Visual inspection of the data, for example by looking at the distribution, is an essential part of the verification process. It can be an efficient way to verify problems after detecting for example outliers or a deviation in distribution. However, once the data has improved in such a way that its form no longer shows any worrisome elements, this should not be used as proof that the data is now correct. At this point in the process, other methods are needed to proof that the data quality is good (enough).

One method is to compare a variable with another variable which is supposed to measure the same thing. In our verification project we for example compared time passed according to the time stamp with time passed according to distance travelled divided by the speed. This led to the discovery that the time stamp was not accurate even though the DtSPAD data looked good upon form inspection. In our case, the time stamp in the dataset was not the actual time logged by the GPS sensor but the time that the logging took place of the GPS signal once it arrived at a server where the time of the server was taken. Due to differing latency times this led to cases in which the timestamp indicated that two seconds had passed while in fact, given the distance travelled and the speed, zero to seven seconds had passed. This varying time deviation was problematic for our indicator because it can lead to relevant data points not being included (still approaching a red light but data no longer included).

While the use of a different timestamp than the GPS time might seem strange, it made sense to the persons who had set up the system. The alternative time was what they called "the train time" and this time made it easy to combine different measurements because they all had the same "train time" and the time latency was not a significant issue for their usage. It just never occurred to them that it might be a problem for the DtSPAD project, just as it did not occur to us before verification that there could be another "time" than the actual (GPS) time.

To further improve verification, Van Gelder and Vrijling [139] highlighted the importance of extending visual inspections and statistical homogeneity tests with physical-based homogeneity tests. By considering whether the data can be split in subsets based on physical characteristics of the individual data points, it can be prevented that the variable as a whole seems homogeneous, while it is in fact a combination of two or more different

distributions that could, by chance, look like one homogeneous distribution when put together.

6.3.2. Pitfall 2

6.3.2.1. Example

For the DtSPAD indicator we created a categorizing variable indicating whether a yellow aspect was planned or not planned. This variable was not always correct. We discovered that sometimes a yellow aspect was characterized as "unplanned" while it was in fact part of a planned arrival. It turned out that short stops of trains were not yet included as planned stops. A bug fix was done to include the short stops. What is now our view on the quality of the indicator?

6.3.2.2. The Improved-Thus-Correct Fallacy

The intuitive response is to think the planned/not-planned indicator is now correct. This is called the "improved-thus-correct" fallacy. In reality, the quality of the indicator is not necessarily good after improvement. The improvement can have caused new problems, especially in coding where bug fixes can create new bugs. However, even if the improvement was implemented correctly, there can still be problems within the data which are not fixed by this specific improvement.

These are straightforward notions, yet we tend to forget them which leads to the improved-thus-correct fallacy. This fallacy can present itself by someone saying an indicator is correct after it has been improved without knowing the actual quality, but more often the fallacy will result in someone not explicitly stating the quality is now good, but forgetting the need to recheck the quality.

This phenomenon can be clarified by thinking of the structure of knowledge in our brain and the automatic activation of associations. Imagine the concepts "Algorithm" and "Improvement" being present in our brains. In the situation as stated by the example, we are aware that something is wrong with the algorithm and thus it is associated with "something is wrong" and not yet with "improvement". Activation of "algorithm" will now also automatically activate "something is wrong", while activation of "improvement" activates other positive concepts like "good" and "solution" (See **Figure 28**).

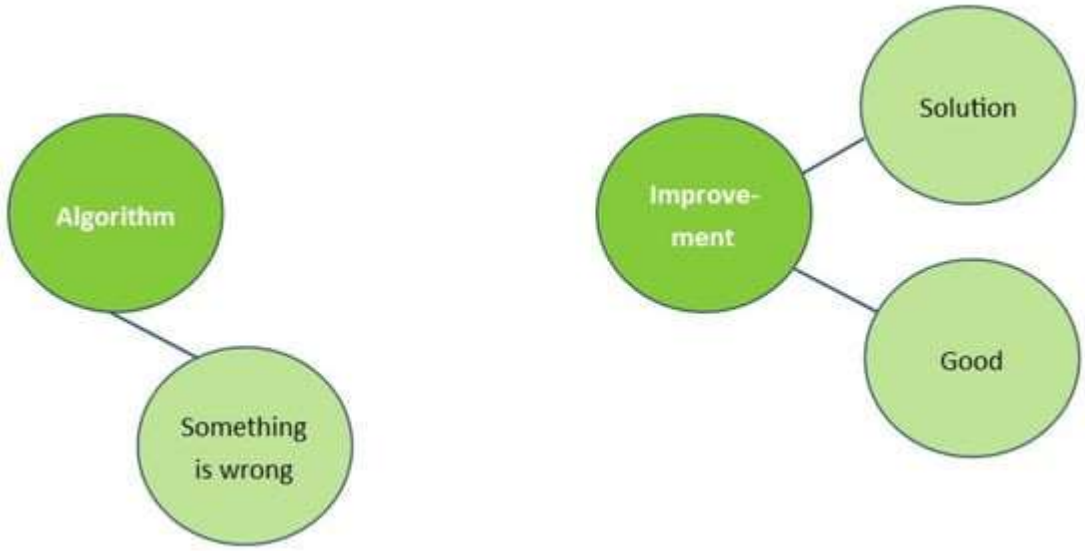


Figure 28. Associations and automatic activation before fix.

After the bug fix, the notion of "something is wrong" changes to "something was wrong" and "algorithm" is now also associated with "short stop was not included" and "improvement", which both share "addition of short stop" (See **Figure 29**).

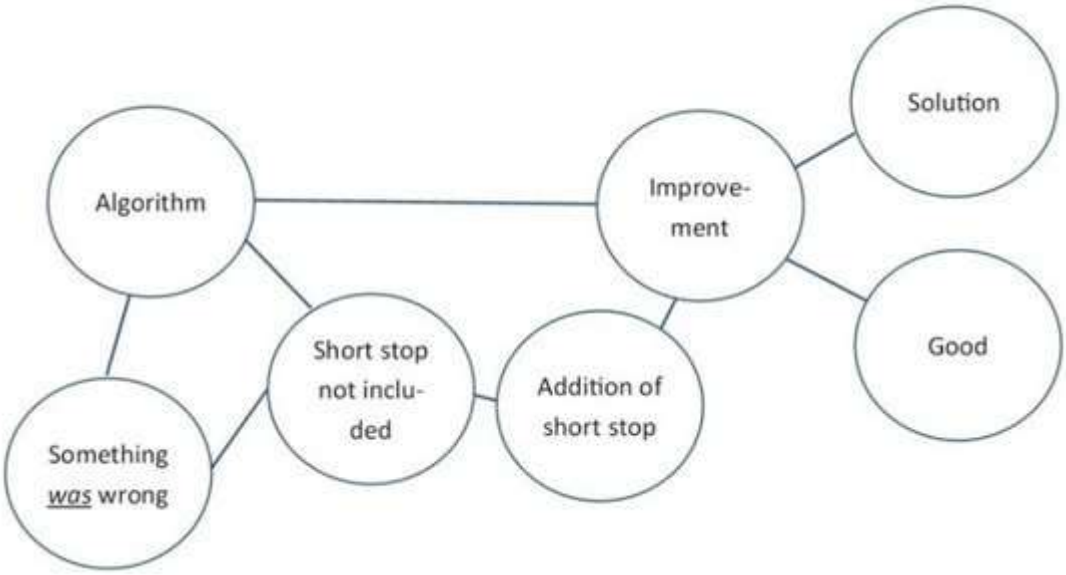


Figure 29. Associations after bug fix.

The activation of "algorithm" will now also activate "improvement" and "good". The aspect "something was wrong" will still be activated as well, but this enhances what was wrong "missing short stop" and then the solution which again is connected to "improvement". At the same time, the idea that there might have been other causes as to why something was wrong with the algorithm is not automatically activated as it is not connected (See **Figure 30**). During the process there was no learning and thus no reason for neurons to connect between "something is wrong" and any other cause which does not have a concrete representation yet, unlike for example "short stop was missing" which can be vividly activated. That is to say, other possible causes "do not have a face" and therefore are not automatically activated while other concepts are, providing a system 1 answer that is easy to accept.

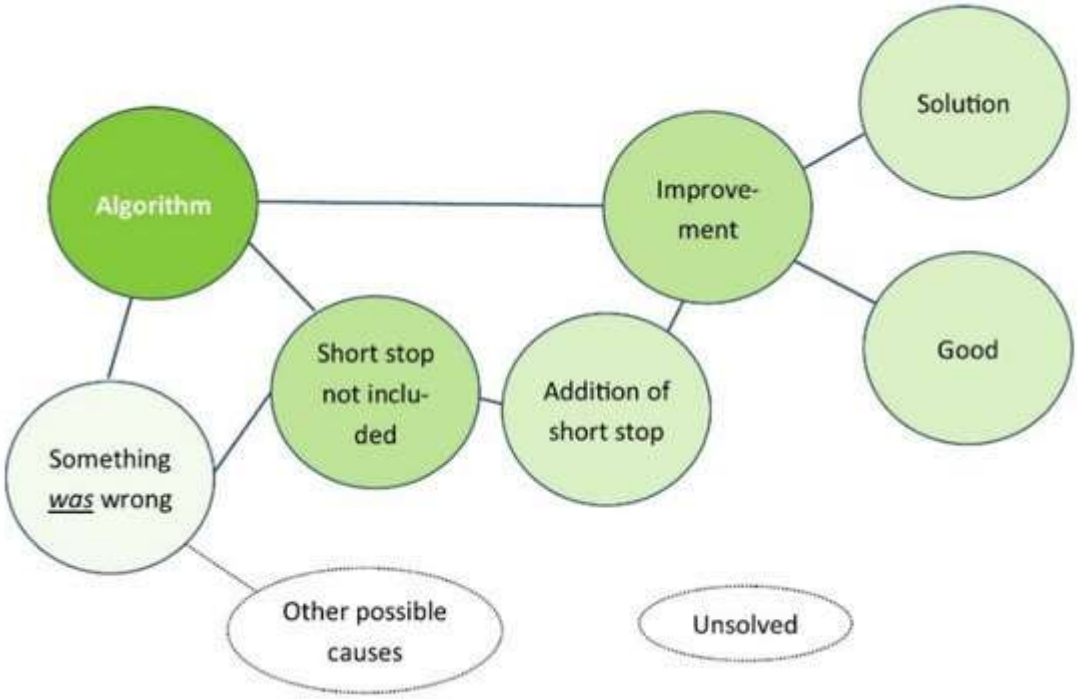


Figure 30. Associations and activation after bug fix.

In our verification project it was noticeable that when the generic status of the planned/unplanned indicator was questioned, the answer was: "there was a bugfix with a high impact two weeks ago to include short stops". While it was then not explicitly said that the indicator was now correct, the effect of the fallacy was noticeable in the fact that we tended to forget to check the current quality of the indicator. Even though it was part of the to-do list, it needed explicit reminding, otherwise it was simply overlooked. Even when the indicator was checked, the implicit assumption was that it would now be correct, noticeable by the sense of surprise when discovering new problems. This sense of surprise also occurred for another indicator which was improved and a new check was done in the sense of "just a formality", which to our surprise exposed the need for more improvement.

6.3.2.3. Implications

This fallacy highlights that people tend to overlook the need to check something again (e.g., an algorithm) after improvements. Therefore, it is important to create an explicit step within the verification process to perform a quality check after every improvement. Additionally, it is important to phrase the current quality not in statements of last improvements, but in a number or unit, like % unknown or % error, or even something more qualitative, like "checking for 5 h did not lead to the discovery of new errors". Even if the current quality cannot yet be specified, the empty field will indicate the need to (re)do a quality check.

6.3.3. Pitfall 3

6.3.3.1. Situation-Dependent-Identity-Oversight

When thinking about the quality of an object, two problems occur. One is that it is hard for us to imagine all factors that can influence the quality. Examples include human factors issues, like things that can go wrong during installation, or the influence of human behavior on the collected results.

Discovering that the quality is very different than expected because of an unforeseen factor is usually followed by the phrase: "I did not think of that". While this can be a serious problem, the inability to think of such factors is not a system 1 problem.

In fact, it is a problem that remains, even when we use our system 2 thinking, since it is more a matter of the knowledge we have, our previous experiences and creativity. Being aware of our inability can help us to collect more information or choose different approaches, like performing verification measurements on the sensor once it has already been installed.

This is however where the actual system 1 problem, the cognitive pitfall comes in: we have the tendency to overlook the fact that objects actually have differing identities or differing qualities in different situations. We do not think in terms of "this object = x in situation A and the same object = y in situation B". Instead we just say "this object is x". For example, when I ask you, what color do the leaves of an oak tree have? Your answer will be "green". Anyone will accept this answer as true. Anyone will agree that indeed the leaves of an oak tree are green. We collectively accept this truth, even though all of us also know that the leaves are not always green. The fact that, even though the oak trees' leaves are orange or yellow or brown in the fall, we still say the leaves are green, provides some inside in the way knowledge is structured in our brains. **Figure 31** shown below illustrates a hypothetical structure. The concept "tree" is linked to many other elements, including "has leaves", which is connected to "except in winter" and to "color green", which is connected to "except autumn" which is connected to "color red/yellow/orange".

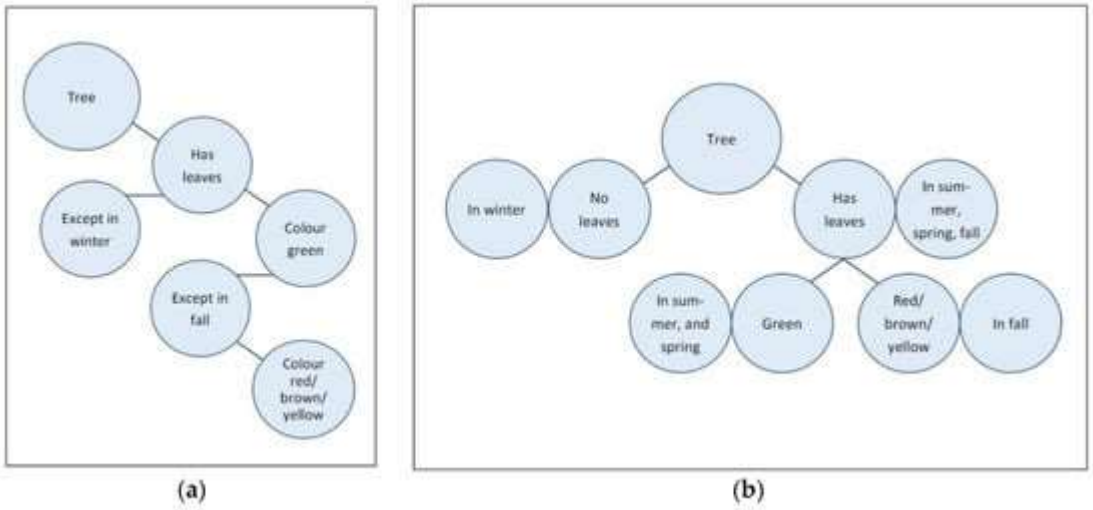


Figure 31. (a) Hypothetical knowledge structure A; **(b)** hypothetical knowledge structure B.

A model that incorporates a situation-dependent-identity would look more like **Figure 31b** above. The model on the right needs a lot more nodes to hold the same information. The structure of the model on the left makes it possible to get to a first answer quickly and efficiently (via automatic activation), with the possibility to obtain the rest of the knowledge when thinking more about it (system 2).

The advantage of the model on the right however is that it is more noticeable when you do not know the answer in a specific situation, for example the color of leaves in fall, since there will be a blank node connected to that situation. In the model on the left, on the other hand, when you do not know the color of the leaves in fall, the bottom of the model will simply fall off and you can still answer the question ”what color do leaves of trees have?” without any empty spaces (see **Figure 32**).

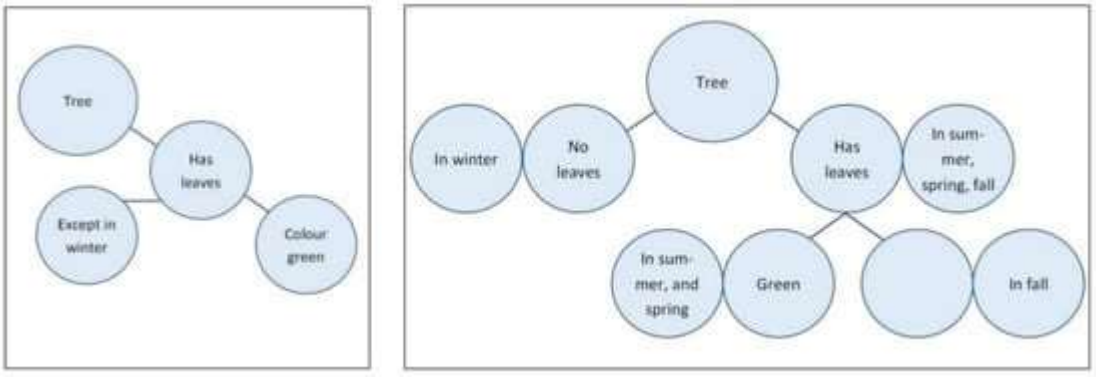


Figure 32. Saliency of missing information in both structures.

6.3.3.2. Example

A hypothetical big data project uses temperature as one of the variables to calculate the final indicator. The used digital temperature sensor has been tested in the lab and logs the temperature every 30 s with an accuracy of 0.3 °C. For this project, an accuracy of ±0.5 °C or better is sufficient. Is the data quality of the sensor sufficient for this project?

The intuitive answer is yes. However, additional relevant questions are: Is the sensor installed correctly and in the correct place when used for the project? Has it been calibrated (repeatedly)? Does it work in the used context? The impact and vibration caused by trains driving over the tracks might disturb measurements after the sensors are installed on tracks. Does the sensor have the same accuracy over the whole range of measured temperatures? Are human acts needed to turn on/off the sensor? Are there any other context factors of the implemented sensor that could impact its accuracy or the logging frequency?

6.3.3.3. Cases

During the DtSPAD project, the tendency to think in terms of one (situation independent) description was for example noticeable with respect to the GPS-sensor. We had seen plots of the GPS locations of a train and noticed that these follow the tracks. When asked about the quality of the GPS sensor, we were therefore inclined to answer: “fairly good based on initial observation”. Sometimes we forgot to include the phrase “but the quality is very bad when the train is located in a train shed or under a platform roof”. As well, we tended to forget the possibility of other factors impacting the quality as part of our check list. Even though these elements can come up when time is devoted to this specific topic and people are in system 2 thinking mode, they might be overlooked at other times, especially during (verbal) handovers to other people or in the interpretation by other people based on written handover.

Another example of this pitfall occurred during the analysis of an error with the previous version of the indicator, the Time to SPAD. This previous version looked at the remaining time available in seconds, before an emergency brake needed to be applied, instead of the required deceleration. In the dataset, we discovered trains with a negative time,

indicating that they would pass a red aspect, followed by a positive time, which should not have been possible. This suggested a problem in the calculation of the minimal braking distance. The minimal braking distance was calculated taking the parameters into account related to the safety brake/quick-acting brake (see **Figure 33**).[8]

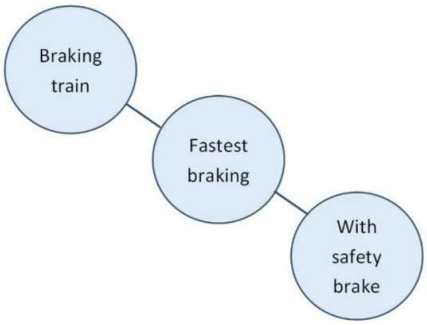


Figure 33. Knowledge structure minimal braking distance.

Repeated checks of the execution of the formula did not lead to any further insight. Eventually it was noticed that using the safety brake only leads to the shortest possible braking distance when the train is driving at a certain speed or faster. At very low speeds, the braking distance is shorter with the regular brake. The conceptual validity of the formula had been checked before as part of the verification but not considered as the problem. Only after other issues with respect to its execution were scrutinized and deemed well, was the conceptual validity checked again as part of the same verification and the problem found. This example again shows that thinking in terms of situation-dependent deviation is not a first intuitive approach, especially since the question ('what is the fastest braking/minimal braking distance') can already be answered ('with safety brake/ this formula') (See **Figure 34**).

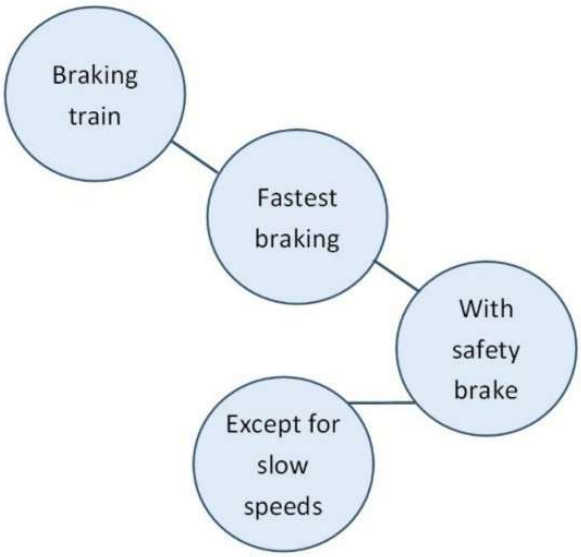


Figure 34. Relevant knowledge comes after the initial answer in the model.

[8] The wording in this paragraph and the next paragraph implies that there are different types of braking systems. This is not the case. The safety brake or quick-acting brake that is referred to here, is in fact a suboptimal translation of the Dutch word 'snelremming' which more literally translate as 'fast braking' and refers to putting the normal brakes on with full force in combination with shorter application times of the brakes thanks to values which can speed up the drop in air pressure in the main brake pipe. I have added this footnote instead of adjusting the main body of the text since the original texts more accurately reflects the thought processes as they occurred at the time and which are the main topic of this chapter.

6.3.3.4. Implication

To prevent oversight of factors influencing the quality, it is important to include in the description of the quantification of the quality (accuracy and loggings frequency) for which situation this quality is applicable. For example, for the temperature sensor it could say: an accuracy of 0.3 °C when installed correctly and calibrated in a lab environment whilst measuring temperatures ranging from –10 °C to 40 °C. For other elements, it can be useful to include details about software package and expected human behavior in operation. Since it is difficult to oversee all possible factors influencing the quality it can be useful to test the quality whilst in the actual operation mode used for the project, check in a number of different ways, like with other software packages or other types of code, and to learn from other projects about which factors had an (unforeseen) effect.

6.3.4. Pitfall 4

6.3.4.1. Example

We calculate a DtSPAD value for a train on every moment a GPS location is logged. What is the influence on the DtSPAD indicator if the logging frequency would be reduced from once per two seconds to once per three seconds? Does this have a problematic impact on the quality of the DtSPAD indicator?

Based on this information, it is very hard to answer the question in detail. An exact answer does not come to mind, but system 1 immediately provides a response like "it is not really a problem". We can however simply not know at this moment, regardless of our hunch that it is not very problematic. For demonstration purposes let's consider the case of lowered logging frequency in more detail.

The top left image in **Figure 35** displays the ideal situation with continuous logging in which three variables are combined to create the indicator. In the situation on the right, some values are not logged leading to lower coverage in the indicator.

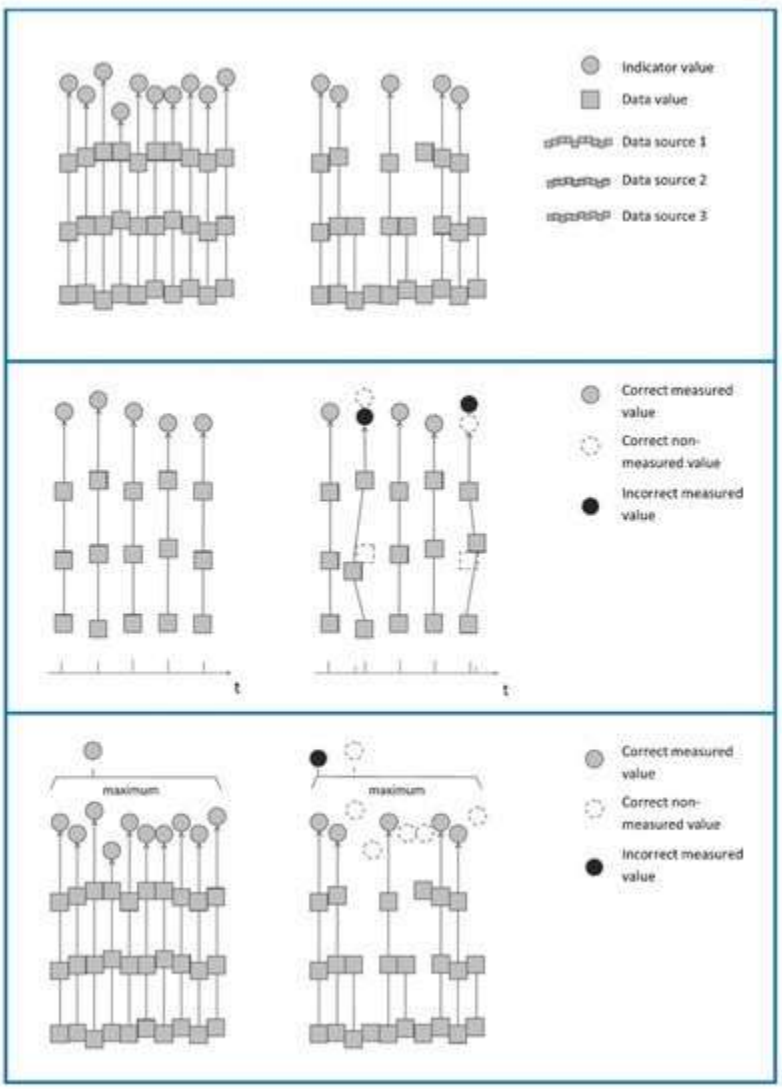


Figure 35. Complicated relation between logging frequency and indicator.

In the second situation (middle pane of **Figure 35**), there is not only lower logging frequency, but because the different variables are not logged at the same time, they need to be matched leading to lower coverage in the indicator and deviation from the actual indicator value.

In the final example (bottom pane **Figure 35**), the actual indicator is a selection from the data, like in our project we are interested in the maximum DtSPAD. In this case, there is deviation from the actual indicator value because a different data point is selected.

6.3.4.2. Impact Underestimation

The above example is meant to illustrate how complex the relation can be between one variable and the indicator and how difficult it is to oversee the impact of changes in a variable on the indicator. When we are confronted with questions like the one in the example, we actually do not know the answer, until we thoroughly analyze it. Yet, such questions do not trigger a sense of panic, because of two reasons (1) even without an

exact answer we still have a rough idea that it probably does not have that much impact and (2) within humans, a sense of danger arises at the presence of signs of dangers and not by the absence of signs of safety. The former reason can be explained again by automatic activation. The small number related to the variable installs the idea of a small effect. This is a manifestation of the anchoring bias in which a specific number influences someone's numerical estimate to an unrelated question [118,121].

6.3.4.3. Implication

The impact underestimation can cause small deviations in quality of a variable to, unwarrantedly, fly under the radar. When the quality of the indicator is critical, it is important to substantiate the impact of the given quality of each variable on the indicator. This requires thorough analysis and system 2 thinking whilst disregarding the system 1 feeling that there probably is not a problem.

6.3.5. Pitfall 5

6.3.5.1. The Beaten Track Disadvantage

When verifying data, texts, theories, code or formula's, some persons prefer to take the "blank slate" approach and first view the to-be-verified element and then judge. This is in contrast with first thinking of your own version of the data/text/theory/code/formula and then comparing it with the to-be-verified element. The beaten track disadvantage entails the tendency to overlook problems during verification when employing the "blank slate" approach. During this approach, reading the to-be-verified element will activate the read concepts in your brain including the logic of the story and this will create a beaten track. This beaten track is easy to travel along in the sense that it is highly activated and the first ideas come up again easily. Alternative notions are activated less easily and can "lose" from the beaten track.

A comparable experience we all have had is during brainstorm sessions, for example when considering a new approach or a project title or a gift for a friend. When a few suggestions have already been made, you have the tendency to think of these same suggestions again and again or the same suggestion in slightly different wording. As soon as you feel you are on the brink of thinking of something new, another person offers their suggestion and you lose your own thought. It cannot compete with the other activation.

The beaten track disadvantage does not necessarily prevent one from noticing erroneous thinking. It mainly prevents you from including elements in your judgment that are not in the to-be-verified object, therefore causing you to overlook certain elements. You are able to judge what is done, but less able to judge what they have not done and should have done. In a verification task it is therefore useful to first consider your own version of the correct solution and then compare it with the actual solution. Even just considering all the factors to take into account and the necessary information to be collected can already be useful.

6.3.5.2. Example

Below is a simplified example of a hypothetical calculation process:

"The possible locations to install the GPS sensor on the train differ per train. Each train has one possible location to install the GPS sensor with a set distance to the head of the train. The FTS train 4 has a distance of 54 m to the head of the train. The FTS train 6 has a distance of 86 m to the head of the train. Of all the FTS trains, only the FTS 6 trains have been equipped with a GPS sensor so far. The distance between the head of the train to the signal is calculated using Vincenty's formula to discover the distance between

longitude and latitude provided by the GPS sensor and longitude and latitude of the signal. The 86 m are then subtracted from that sensor-signal distance to get the distance between the head of the train and the signal (accepting some inaccuracies due to track curvature instead of a straight line between the train and the signal)."

For this simplified example, a question during the verification process would be whether there are any issues with this process as described above?

When we accept the usage of Vincenty's formula in this case, then there does not seem to be a problem with the process. Yet there is one potentially relevant question to ask: do we also receive GPS data from trains other than FTS trains? If so, then the adjustment for GPS sensor position might be incorrect. This seems like a straightforward question to ask and yet it is easily overlooked. The presence of more trains than only FTS trains with GPS data is a realistic situation, especially when different parts of the algorithms are written by different persons and they receive the information in fragments. The programmer who includes the distance from the sensor to the head of the train might be referred to someone who is knowledgeable about these distances. When this person only ever works with FTS trains, he or she will only give the information related to the FTS trains and the programmer will use this knowledge in his coding.

6.3.5.3. Cases

One of the cases that occurred during our verification project did entail the difference between sensor location and head of the train. During the calculation of the DtSPAD, the calculation of the distance between head of the train and the red aspect took the sensor location into account. However, during an earlier version, calibration of the time was also necessary because the clocks of the sensors were not running synchronous. The calibration was done by taking the moment in time according to the GPS sensor when they passed a signal and the time according to the system in the tracks registering train passage.

When we looked at the code written for this calibration, we did not register it as a problem that the actual longitude and latitude of the GPS sensor were used instead of the adjusted location of the head of the train, even though we were familiar with the issue of sensor distance even for calibration in other settings. However, when looking at the code, which was otherwise executed perfectly fine, the problem was not noticed. During code adjustments for the calculation of the DtSPAD with respect to the sensor distance, the programmer noted he should use this approach for the calibration as well upon which the response was: "did you not already do that?", illustrating that the knowledge of its necessity was there but it was not sufficient for us to notice the glitch when looking at the code.

6.3.5.4. Implication

If quantitative verification with actual data is possible, this is a sound approach. When theories or algorithms need to be checked, this is however not always possible. In these types of expert-judgment verifications it is useful to discourage the "blank slate approach" and encourage persons to first consider their own version of the correct solution and then compare it with the actual solution. When this is too time consuming, one can restrict the work to considering the factors to take into account when one would try to create the correct solutions themselves. These factors can then be used as a checklist or backbone to verify the element with.

6.3.6. Summary of All 5 Identified Pitfalls

The five identified pitfalls are summarized in **Table 10** with a recommendation per pitfall.

#	Pitfall Name	Description	Recommendation
1	The good form as evidence-error	The incorrect assumption that if data looks good, for example in terms of distribution, that the quality is therefore good.	Starting with form checks is important, but make sure to check in other systematic ways as well by for example comparing sources that are supposed to measure the same variable.
2	The improved-thus-correct fallacy	The incorrect assumption that if the data is improved, for example because of a bug fix, that the data quality is then good, or more subtly, forgetting to recheck whether the data is actually good after the improvement.	Develop a procedure to recheck the data after every new improvement and express the data quality in terms of actual quality instead of bugfixes. Keeping a list of the quality of each variable at certain dates can be useful.
3	Situation-dependent-identity-oversight	The tendency to forget that data, for example coming from a sensor, can be of different quality depending on the situation.	When writing down the quality of a variable/data source, include a description of the condition in which this quality applies (especially when applies to lab tests versus in position). If unknown, leave a question mark to visualize that the listed quality might not apply in other circumstances.
4	Impact underestimation	The incorrect assumption that small variation in a data source corresponds with small variation in the outcome.	When the outcome is critical, assume that it is impossible to grasp the impact of a variable unless studied and simulated explicitly. Keep track of the decision which variations are accepted and which are not.
5	The beaten track disadvantage	The difficulty to spot problems when following the narrative of the to-be-verified item.	Use systematic verification where possible. If expert judgement is necessary, make sure the expert forms an opinion before verifying the to-be-verified item.
Generic recommendation			
ALL	Create awareness regarding system 1 thinking, mainly focusing on the fact that data verification is complicated and (big) data projects include complex interactions. Solutions/conclusions that come to mind easily are likely based on system 1 thinking. Given the complexity of the tasks at hand, it is possible that these solutions/conclusions are not based on all relevant information and/or include implicit incorrect assumptions that work in general in life but not with respect to (big) data. Teams are important to help each other to think of and consider all the relevant information and to set aside time to reconsider previously drawn conclusions.		

Table 10. Summary of the 5 identified pitfalls.

6.4. Discussion

6.4.1. Limitations and Further Research

This "proof of concept" case study of our mDtSPAD safety indicator took a closer look at the human factors challenges in the verification process as part of (big) data utilization. Five cognitive pitfalls were identified to be aware of when verifying data, given the way our brains function. It is expected that knowledge of these pitfalls is relevant for other railway organizations and other industries as well, because cognitive biases in general have been proven to occur amongst all people. However, the prevalence of these pitfalls and data verification within other organizations is not known. The current study focused on testing a more extensive theoretical framework on cognitive biases in an actual setting, with a focus on providing a deeper understanding of these types of errors and their prevention. Future studies that focus on measuring the prevalence of these pitfalls would be beneficial, followed by research on the success rate of interventions.

Another limitation of the current research is that the list of five pitfalls is not necessarily exhaustive. It is possible that there are other cognitive pitfalls relevant for the verification process that are not in this list because they did not occur during this specific case study or did not lead to salient errors. Further research to identify other possible cognitive pitfalls can consist of other case studies or experimental settings with respect to data verification. This research is especially important in use cases where the results cannot be easily verified, that is when the calculated indicator does not have an equivalent indicator or predicted live data to compare it with. This is the case for safety indicators that relate to low incidence incidents, like SPADs, but can also be the case for "softer" measures, like "safe driving behavior" or, for example in health care, for measures like "improved health" or "surgery success".

Besides improving the verification process, future studies are also needed to improve other aspects of (big) data: Even when the input data is correct, the results can still be incorrect. Common errors include the sampling error causing the data to be non-representative. Even in the big data domain where the assumption often is that we have all the data, this can be a far cry from the truth if there are non-random gaps in the data [114]. Multiple comparison is also highlighted as a big data issue, meaning that the presence of a lot of variables and a lot of data will, by chance, always lead to some seemingly significant factors unless corrected for. Van Gelder and Nijs [140] also note this issue in their overview of typical statistical flaws and errors that they found upon investigation of published big data studies related to pharmacotherapy selection.

Another problem is that big data solutions are notorious for focusing on correlation and ignoring causality. The Google Flu prediction algorithm for example was based on the amount of flu related google searches and was considered an exemplary use of big data, until the predictions were far off in 2013. The overestimation was likely caused, at least in part, by a media frenzy on flu in 2013 leading to a lot of flu-related searches by healthy people. Additionally, the constant improvements in Google's search algorithm has likely had an effect on the quality of the predictions [141]. Even in cases where the prediction model can be updated based on new information of the changing circumstances, this might already have led to losses when the results were acted upon and the cost of a false-alarm or miss are high. The universal occurrence and especially recurrence of such errors (e.g., not taking changing relationships into account, multiple comparison and sampling error) can be illuminated by investigating the role that cognitive biases play in their occurrence.

The interpretation of data and the data results can also be complicated. This is a vast topic on its own that deserve many references. Within railways specifically, Figueres-Esteban and colleagues (2015) discuss the challenges of comprehending big data within and advocate a decision support system specifically for the railways. See [142] for their recommendations and literature review.

6.4.2. Advocated Perspective and Recommendations

For each cognitive pitfall identified in this study, recommendations were given to prevent them from leading to errors. When thinking about tackling errors within risk monitoring, it is important to keep in mind that, given the way our brains function, it is expected that errors occur within information judgment tasks as part of risk monitoring. These errors occur regardless of the intelligence of the persons performing the tasks and are not person-dependent. Creating awareness among persons about these facts, the way we think and our tendency to fall into these pitfalls is part of the approach against cognitive pitfalls. Secondly, and equally if not more important, measures can be taken to improve the process itself and minimize the chances that people will make these types of errors.

This second approach consists of formalizing the verification process to create reminders of the factors that need to be considered with system 2 thinking so the errors do not occur.

These formalizations of the verification process are not designed to take away some of the cognitive load or the thinking of the persons involved, but in contrast to encourage deep and reasoned thinking. It is a matter of setting up the right circumstances, of facilitating the possibility of persons to be able to handle the cognitive task in the desired manner: with system 2 thinking and thereby their own, well-based, judgment.

Although this chapter might appear to highlight human's limitations, it is actually meant to illustrate that there are many instances in which we do not use the full extent of our capabilities which causes errors rather than a lack of capabilities as a cause of these errors. With the right adjustments in processes and increased awareness, we do not become more intelligent, but we are able to perform as if we did. Especially in the age of (big) data usage where information judgment takes on a new level of complexity, while also providing unique opportunities to improve safety, facilitating the best possible performance of the human brain via work process improvement is not a matter of optimization but of necessity.

This study and other referenced examples make it apparent that we tend to have false assumptions: we implicitly assume that when we look at data, it is correct or we would notice and that persons looking at the data before us would have noticed if anything was wrong. As the use of (big) data is becoming more common, it is becoming increasingly important to tackle these issues. If we want correct conclusions, we need good quality data and if we want good quality data, we need to set up a solid verification process befitting human cognition.

This case study has also shown that, in the larger conversation of improving data utilization, considering technical advancements alone is not enough: a focus on the human factor in the (verification) process is essential to truly fulfill the grand promises of big, and medium-sized, data.

Chapter 7

Practical insights on using big data to investigate human behavior and improve safety

7.1. Using big data: Lessons learned for behavior related safety research and process improvements

The use of big data for this dissertation provided multiple insights that are useful for future endeavors using big data for behavior related safety research within academia, rail and other industries.

7.1.1. Beyond statistically significant: Using big data to identify effect size and exact circumstances

The last decades of research within human factors and other disciplines have shown that many factors can influence human behavior. Building upon this knowledge, the question is shifting from "can a factor influence behavior" into the question: "to which degree does a factor influence behavior?"

Step one has in large part been done within science: the identification of potential factors. Existing theory and experts in the field help us to identify and operationalize the factors. Step two is the ordering of these factors in effect size and investigating under which exact circumstances the effect occurs. The use of big data takes us outside artificial laboratory settings and provides knowledge on the effect size in real-life settings that occur day-to-day.

Depending on the dataset being used, big data can easily lead to statistically significant results. In the case of results that are statistically significant, the focus lies on how big the effect is. Very small effects or even non-significant effects (despite large amounts of data) also provide valuable knowledge as it is often also informative to know what does nót have a (substantial) effect when it comes to practical application.

7.1.2. The human error in big data research

Any task involving human performance is susceptible to human error, including tasks such as data analysis. Those that investigate human behavior are expected to have a certain level of knowledge about human behavior and human error. We should take advantage of this knowledge and not be afraid to apply it to ourselves and our processes. No one is immune to being human.

Using big data provides some specific openings for errors to occur and remain undetected. Firstly, due to the amount of data, individual inspection of all the data points is unlikely. Secondly, the data might be originally collected for other purposes and have hidden or uncommunicated properties that serve the original goal but not the current research goal. Thirdly, multiple variables might be combined leading to many potential entry points for problems in data quality.

One might be aware of some unknown information, but there will also undoubtedly be unknown unknowns. Thorough interaction with the data in different manners, like examining specific data points, will help to get a better grasp of the data and the unknown unknowns. For concrete advice, see chapter 6.

7.1.3. User-friendly data: the importance of a useful indicator

The indicator described in this dissertation is the "maximum deceleration-to-SPAD". It summarizes each red aspect approach into one number. This indicator was not the original indicator. At the start of my research, the indicator that had just been developed and taken into use was called "minimum Time-to-SPAD". Being able to summarize all the

data points of an approach into one relevant number was an important first step, but not enough.

The original indicator made intuitive sense. For each data-point (train speed and distance to red aspect), the time was calculated (in seconds) until the train had to decelerate with the emergency brake in order to stop exactly at the red aspect. This calculation included the time that it takes for the emergency brakes to apply. For each approach, the minimum time was selected. If the minimum Time-to-SPAD was, for example, ten seconds, then this meant that the train driver had ten seconds left until he or she had to use the emergency brake in order to prevent a SPAD. If the Time-to-SPAD was one second, then it was a close call.

One of the difficulties with the minimum Time-to-SPAD, was identifying the cut-off score. Is ten seconds low? Or five or only two? While identifying a cut-off score is often difficult, examination of individual approaches with different minimum Time-to-SPAD values, led to different conclusions from different subject matter experts. There were only a few approaches where there was consensus amongst subject matter experts. Additionally, the minimum Time-to-SPAD value did not seem to provide sufficient information to the experts since they also wanted to know the train speed at the moment of the minimum Time-to-SPAD. A final clue was that most subject matter experts wanted to see how fast the train had decelerated before they could provide a conclusion.

It became clear that the actual deceleration of the train mattered to the subject matter experts. Solely looking at the actual deceleration was however also inherently flawed as an indicator for the probability of a SPAD. A train driver might decelerate very fast at a long distance in front of the red aspect and drive very safely with respect to SPAD probability. In such situations, the fast deceleration was not actually necessary to be able to stop in front of the red aspect. These notions eventually led me to consider: what if we calculate the deceleration that is actually necessary?

Upgrading the indicator from time-to-SPAD to deceleration-to-SPAD had many advantages:

- Fewer estimated fixed variables were needed whilst no additional data point were necessary to calculate the maximum Deceleration-to-SPAD.
- The indicator value is easier for subject matter experts to interpret and to judge whether a value is low or high.
- There was more consensus amongst subject matter experts about high-risk approaches. One specific considerable improvement was that low-speed approaches were no longer incorrectly labeled as “near miss”.
- The new indicator had a normal distribution in contrast to the original indicator which had no clear distribution shape with values as high as hundreds of seconds. Having a variable that follows the normal distribution or a comparable distribution has many advantages from an analysis perspective. Some of these advantages have been discussed in chapter 5.

7.1.4. The importance of a theory-driven approach in safety context

There are many ways to analyze big data. One can use modern techniques that rely less on the traditional "top-down" theory-driven approach and focus more on a "bottom-up" approach whereby the algorithms help select the best combination of factors for a predictive model. Whilst these techniques can provide very useful insights, they are not necessarily useful for certain safety-related research.

Two important factors to consider are the cost of a "false alarm" and, related, whether there is a need for a causal explanation. At ProRail, a predictive model has been developed which predicts locations where people will walk along the tracks. Inspectors can use this information to inspect a specific location if they happen to be in the area. When the algorithm made the right prediction, this is great, and when the algorithm was incorrect, then there is no real loss. However, when an algorithm predicts that there are certain locations with a high SPAD risk, then the question becomes: what now? Adding additional (technical) safety barriers in Dutch rail tends to cost a lot of money. The results from an algorithm with for example a 67% predictive performance will not be considered sufficient evidence to warrant the investment.

In the example of high SPAD locations, it is also valuable to know why there is a risk, so the cause can be eliminated if possible rather than adding additional safety barriers. Whilst the algorithms can identify which factors have the highest predictive performance, there might be spurious correlations present. If this is the case, changes to the factor might have no or even a worsening impact on the risk that one tries to reduce. A theory-driven approach can help in providing a higher degree of certainty and a clearer indication of causality.

Another advantage of a theory-driven approach is that certain factors can be investigated that would not have been identified by the algorithms. For our research on incidental learning, one of the additional variables that needed to be calculated, was the frequency of the same yellow+number aspect in the past 14 days for the same specific signal, if there was a yellow aspect that was near a station platform stop. This extremely specific combination of variables would not have been found by the algorithm.

The modern "bottom-up" techniques can be alluring due to the idea of requiring less subjective decision making by the researchers or analysist. However, in many ways, this is an illusion. Even with the use of algorithms, many decisions need to be made including what methods exactly will be used and how the data is split in training and test set. Perhaps most importantly, the decision of which factors to include remains present. It is neither practically feasible to gather and connect every possible variable nor is it good for the predictive performance of the algorithm to add every possible variable. A selection of variables must be made, and the inclusion or exclusion of specific variables can lead to an entirely different answer to the same question. Here, again, the unknown unknowns are the biggest risk. Interaction with, and in-depth knowledge of what the data represents helps in making sure the relevant questions are asked and the relevant variables are included.

Figure 36 gives an example of how a psychologist and big data expert can collaborate to answer safety related human behavior questions.

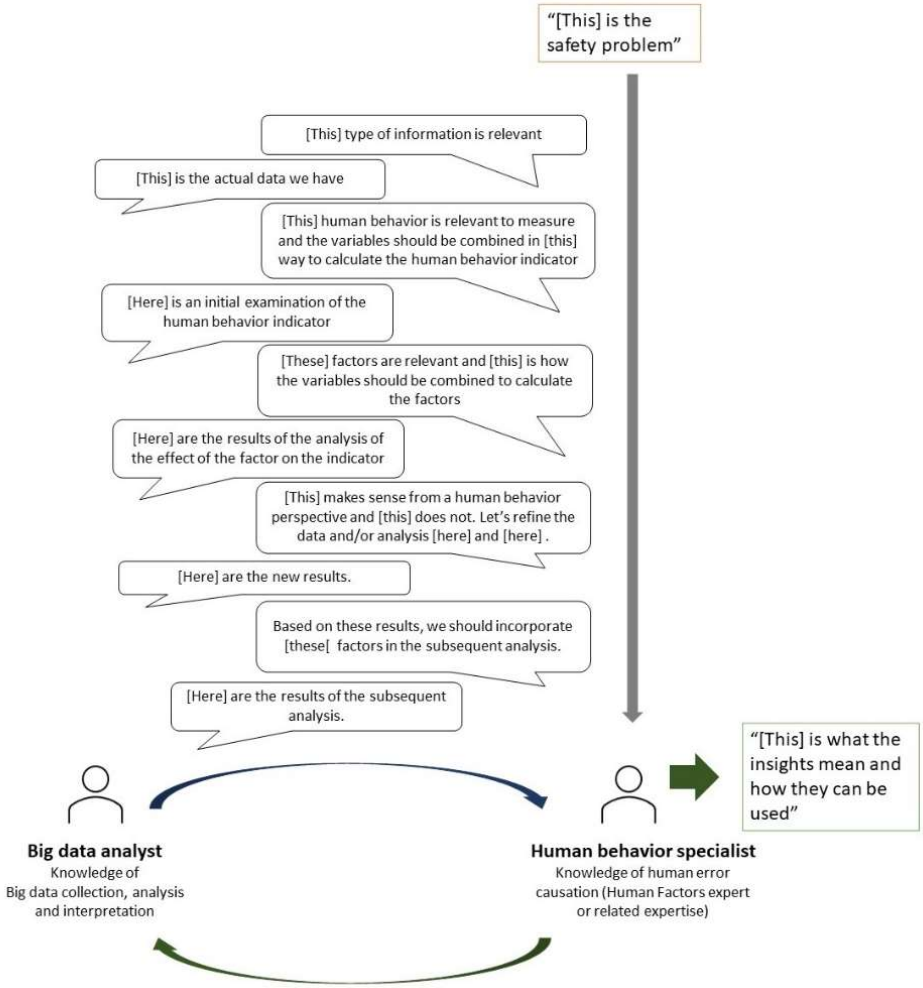


Figure 36. Example of collaboration between Human Behavior specialist and Big data analyst and the added value that the Human Behavior specialist can bring to the process. Both parties need each other to effectively use big data.

7.1.5. Why behavioral data can lead to different findings than accident data

In chapter 5 the philosophy was introduced of looking at variation in processes. With respect to incidental learning, we looked at the variation in train driver behavior. A larger variation theoretically leads to a distribution with more datapoints in the "near miss" region and thereby a larger probability of accidents to occur. However, as we have seen in chapter 3, this is not necessarily the case if there are additional safety barriers in place to prevent the unwanted behavior from leading to an accident.

This means that research using behavioral data cannot necessarily be validated (easily) by research using accident data. This is especially the case for infrequent accidents. Nonetheless, valuable insights can be gained from using behavioral data and from investigating variation, as discussed in chapter 5.

Data variation can also be used as a proactive or leading indicator upon which can be acted before the accident occurs rather than afterwards. Monitoring variation can be a useful leading indicator in general and specifically to monitor the effects of a process change with unknown effects on employee behavior. Within Dutch rail, the leading indicator maximum Deceleration-to-SPAD was used to monitor the effects of an infrastructure change on train driver behavior and the SPAD risk. There was sufficient

data to reliably monitor the mDtSPAD distribution on a weekly basis for passenger trains and be able to conclude whether an intervention was necessary or not. The number of SPADs and the number of high mDtSPAD values was also monitored, but the absence of both provided less (leading) information than the results showing stability in the distribution of train driver behavior.

7.2. Recommendations for safety research using (big) data of employee behavior

Chapter 5 describes the approach of how to examine variation in order to gain new insights and section 7.2 describes tips that are useful for the practical execution. Before any data can be used, there of course needs to be an willingness from industry and/or academia to invest time and money in the endeavor. For this final section we will therefore recommend some arguments that can be used to list the advantages of using (big) data of employee behavior for safety purposes:

- "The methods that we used up until now have led to a reduction in the numbers of accidents, but there still remain accidents which were unforeseen and for which we do not fully understand the cause based on the current methods. Using the data-based approach can provide new insights into the causes of the accidents, and thereby make it possible to:
    - improve accident analysis
    - improve task evaluation
    - develop new or adjusted safety measures for accident prevention and mitigation
    - avoid the unintended introduction of unsafe situations as a result of process changes or innovation"
- "There is an increasing number of a certain accident. The current methods insufficiently explain why this is happening. The data-based approach can provide more insights into the cause of the increase. At the very least, the data can show whether the increase correlates with an increase in opportunity for accidents to occur (exposure) or a change in behavior (chance) or both."
- "The data-based approach allows us to proactively monitor the effects of innovations or changes and intervene before accidents occur. The monitoring based on employee behavior also gives more confidence in the conclusion that a change does not have a negative effect if no to little change in behavior is measured over a time period in contrast to measuring zero accident in the same time period for a location where zero accidents is the baseline. Thus, more firm conclusion can be drawn sooner."
- "There are assumptions within the industry about whether certain factors influence employee behavior and increase the probability of an error, but there is no definitive evidence yet whether this is indeed the case. Even more importantly, we do not know how big the effect is per factor and in which situations this effect occurs. The data-based approach allows us to measure whether the factors indeed increase the probability of an error and which factors have the highest impact."
- "We implement certain costly safety barriers in terms of money or process costs like time, but we have little evidence which locations or situations actually are at high risk for errors to occur. The data-based approach will make it possible to identify high risk locations with more certainty and potentially also identify locations which have a low risk for error if research is done on risk factors. This information makes it possible to make more deliberate decisions on where to implement costly safety barriers and where not to."

Closing remarks

The findings on incidental learning highlight a new and significant factor for risk management and accident prevention. As a cognitive psychologist, I have enjoyed figuring out and elaborating on the theoretical underpinning of these sections. I have made an explicit effort to clearly explain the psychology so readers can go beyond merely obtaining knowledge and into understanding. If this dissertation inspires anyone to continue research on the role of incidental learning in human error occurrence or to change task design in order to prevent errors by taking previous exposure into account, then I will call this dissertation a success.

To continue increasing the level of safety for employees and for society, we need improvements with respect to incidental learning but of course also with respect to other factors that influence human behavior. I believe that the use of big data can lead to great strides in the broad field of behavior related safety issues, both in terms of scientific knowledge and in the prevention of accidents. If this dissertation inspires and supports anyone in this endeavor, then I will call that success, as well.

A final ambition of mine, with this dissertation and in general, is to contribute to fostering more understanding around human behavior. Still too often, we judge and punish others ánd ourselves for the mistakes we make. But when we enter the realm of "he should" or "she should" or "I should have", then it is hard for learning to occur.

Only by accepting what behavior apparently occurs under the current circumstances, can we imagine solutions to change it. This is what I love about using big data from actual day-to-day behavior: it is truth being presented to us. It is not the scientist's job to judge the behavior he or she is shown, but only to explain what is going on and present the results in a way that turns truth into knowledge. Hopefully this knowledge will foster understanding and with it, both compassion towards the persons who have erred in the past and action to prevent future errors and improve safety for all.

"It's not 'us versus them' or even 'us on behalf of them.' For a design thinker it has to be 'us with them'"

– Tim Brown, CEO and President of IDEO

Acknowledgements

I am grateful for a lot of people and thankful for many things that they have done and for who they are. It would take too much space and time to list every person I am grateful for and why, so I am going to try to be brief and not even thank Frank Guldenmund for his support, enthusiasm and valuable insights.

I would like to thank my promotor Pieter van Gelder and daily supervisor Simone Sillem from the TU Delft for their patience and the leeway they occasionally gave me whenever I wanted to do things differently (again) or was simply being stubborn.

I would like to thank many people at ProRail. There are those who silently put in the work to make sure I would be able to start my PhD at ProRail and continue with it. And to every colleague who made it easier for me to go to work at the times that it was hardest to go to work: thank you. Never underestimate the impact you make by simply being there. I would like to thank Jelle van Luipen specifically. You were a fan of me and my work, even when I felt like giving up. That meant a lot. Thank you for always being in my corner.

To my friends and family, if your name needs to be listed here in order for you to know that I love you and am very happy to have you in my life, then please give me a soft hit on the head and I will rectify the situation as soon as possible. Nonetheless, a special shout-out to my parents. I know that no matter what happens, I can always travel back to you and I will be fed and hugged. There is no greater solace than that.

And of course a special shout-out to my friend Mariska, my paranymph. As babies, we were neighbors and occasionally lay side by side in the crib. Thirty years later, I have been your paranymph and you have been mine. Your friendship is a blessing.

And to my other paranymph, Wilco: you are one of those colleagues that is in the friends section. You have been there since day one and have become my safe haven in the Inktpot. I have missed being able to walk over for a chat or a question during the pandemic, but I am comforted by the notion that you are in my life and I am in yours.

Adam Foster. Thank you.

And finally: Jop Groeneweg. When I graduated from Leiden University, you said I should do a PhD and I replied: “I will never do a PhD”. Not even two years later, I started on this PhD adventure with you as my promotor. This PhD would not have started without you and I probably would not have finished it without you. You have been my intellectual sparring partner, my moral compass and the force that kept on pushing me to be even better. I have occasionally found you annoying (let’s not forget how you tricked me into giving a talk for reluctant master students), but you have always been the best mentor and companion I could have asked for on this adventure. May the variation be with you.

p.s. Paul: Thank you.

Curriculum Vitae

Date of birth: July 24th, 1991

Place of birth: Beesel, the Netherlands

Education

2018-2021	Graduate school at Delft University of Technology
2012-2013	Master of Science in Applied Cognitive Psychology at Leiden University. *Cum laude.* Thesis title: "Cognitive Biases and Other Errors in Incident Analyses: An Examination of Tripod Beta Reports to Improve the Learning from Incidents"
2012-2013	Master Honours programme "Leiden Leadership Programme" at Leiden University, 2012-2013.
2009-2012	Bachelor of Science in Psychology at Leiden University, 2009-2012. *Cum laude*. Honours thesis title: "NHST versus PR: A comparison between statistical paradigms based on performance within social psychology"
2010-2012	Bachelor Honours programme "Science & Society" at Leiden University

Employment

2021-present	External PhD student at Safety department at ProRail and Safety and Security Science group, TBM, TU Delft, February 2021-present.
2017-2021	External PhD student at Innovation department at ProRail and Safety and Security Science group, TBM, TU Delft.
2014-2016	Trainer, consultant and product manager at CGE Risk Management Solutions
2014	Junior Human Factors Consultant at Intergo.
2013-2014	Teacher working groups for first year statistics Psychology BSc students at Leiden University

Publications

Journal papers

Burggraaf J, Groeneweg J, Sillem S, van Gelder P. What Employees Do Today Because of Their Experience Yesterday: Previous exposure to yellow+number aspects as a cause for SPAD incidents. Journal of Rail Transport Planning and Management. 2022; 23. https://doi.org/10.1016/j.jrtpm.2022.100332

Burggraaf J, Groeneweg J, Sillem S, van Gelder P. What Employees Do Today Because of Their Experience Yesterday: How Incidental Learning Influences Train Driver Behavior and Safety Margins (A Big Data Analysis). Safety. 2021; 7(1):2. https://doi.org/10.3390/safety7010002

Burggraaf J, Groeneweg J, Sillem S, van Gelder P. How Cognitive Biases Influence the Data Verification of Safety Indicators: A Case Study in Rail. Safety. 2019; 5(4):69. https://doi.org/10.3390/safety5040069

Conference papers

Burggraaf, J.; Groeneweg, J. Je leert meer dan je weet. In Proceedings of the NVVK conference 2019: NVVK info, 23-26.

Burggraaf J.; Groeneweg J. (2017), Cognitieve valkuilen in de verificatie van big data, In Proceedings of the NVVK conference 2017: NVVK Info, 38-40.

Burggraaf, J.; Groeneweg, J. Managing the Human Factor in the Incident Investigation Process. In Proceedings of the SPE International Conference and Exhibition on Health, Safety, Security, Environment, and Social Responsibility; Society of Petroleum Engineers, Stavanger, Norway, April 2016. https://doi.org/10.2118/179207-MS

Burggraaf, J.; Groeneweg, J. Het voorkomen van cognitieve biasen in ongevalsanalyse. In Proceedings of the NVVK conference 2015: NVVK Info, 16-19.

Other communications

Burggraaf, J. Onzekerheid omarmen. Column for NVRB (Nederlandse Vereniging voor Risicoanalyse en Bedrijfszekerheid), 22 oktober 2019, https://www.nvrb.nl/nieuws/column-de-pen/oktober-2019-onzekerheid-omarmen

Presentations and workshops at professional meetings (excluding presentations at ProRail)

Burggraaf, J.M. *Human Factors vraagstukken beantwoorden met realisatiedata: Tips, voorbeelden en waarom HF input nodig is*. 2021, Dutch Rail Human Factors conference, 19 May, online.

Burggraaf, J.M. *Impliciet leren als oorzaak voor STS-passages en fouten in het algemeen.* 2019. Dutch Rail Human Factors conference, 26 November, Amersfoort, the Netherlands.

Burggraaf, J.M. *Je leert meer dan je weet: Invloed van onbewuste processen op handelen.* 2019. Workshop organized for NVVK members via the Andrew Hale scholarship, 14 November, Utrecht, the Netherlands.

Burggraaf, J.M. *Beheerst en veilig? Leer van de spreiding!* 2019. TU Delft MOSHE course, 9 September, The Hague, The Netherlands.

Burggraaf, J.M. *Ook cheeta's lopen wel eens langzaam: Mijn kijk op vakmanschap als wetenschapper*. 2019. NVVK Praktijkdag voor veiligheidskundigen over Vakmanschap. 17 May, Woerden, the Netherlands.

Burggraaf, J.M. *De denkreflex game: Denkreflexen ervaren en leren navigeren*. 2019. Dutch rail conference Risicobesluitvorming: Is er over nagedacht?, 21 March, Railcenter, Amersfoort, the Netherlands.

Burggraaf, J.M. *Je leert meer dan je weet*. 2019. NVVK conference: 2025 Wat ga ik anders doen?, 13-14 March, Arnhem, the Netherlands.

Burggraaf, J.M. *Tunnelvisie in communicatie*. 2018. Veiligheidsbijeenkomst spoorsector Tunnelvisie: Heb jij grip op je veiligheidsblik?, 30 October, Leiderdorp, the Netherlands.

Burggraaf, J.M*. Innovatieve aanpakken en cognitieve valkuilen*. 2018. Netwerkmiddag Veilig werken, Rijksinstituut voor Volksgezondheid en Milieu (RIVM), May 25th, Bilthoven, the Netherlands.

Burggraaf, J.M. *Managing the Human Factor in the Incident Investigation Process*. 2016. SPE International Conference and Exhibition on Health, Safety, Security, Environment, and Social Responsibility, 11-13 April, Stavanger, Norway.

Burggraaf, J.M. *Het voorkomen van cognitieve biasen in ongevalsanalyse. 2015.* NVVK conference: Veiligheid? Zoek het ff zelf uit!, 31 March – 1 April, Arnhem, the Netherlands.

Academic awards

Andrew Hale scholarship at the NVVK conference "2025 Wat ga ik anders doen? Nieuwe risico's, nieuwe opvattingen, nieuwe oplossingen", March 13, 2019, Arnhem, the Netherlands

Appendix A

R code for statistical test:

```
## Statistical testing: get p-value for 1 segment by comparing observed with expected

# example values in total sample

high<-1+39

tot<-38+377

low<-tot-high

values<-c(rep(1,high), rep(0,low))

# example values in subset

high_subset<-1

n_subset<-38

# Prepare for 100000 runs

reps<-100000

result<-logical(length=reps)

# Check side: greater or lesser than; if greater than expected:

if(high_subset>(high/tot*n_subset)){

for(i in 1:reps){

# draw random sample of subset size without replacement and check if as many or more values in drawn as in measured

result[i]<-sum(sample(values, n_subset, replace = F)==1)>=high_subset

}

}

# Check side: greater or lesser than: if lesser than expected:

if(high_subset<(high/tot*n_subset)){

for(i in 1:reps){

# draw random sample of subset size without replacement and check if as little or less values in drawn as in measured

result[i]<-sum(sample(values, n_subset, replace = F)==1)<=high_subset

}

}

p<-sum(result)/reps

p
```

Appendix B

R code for statistical test:

```
## Statistical testing: get p-value for 1 segment by comparing observed with expected
# below values for frequency bin 351-450 times in Table 7.
SPADs<-13+5+0+2+0+5+4+0
RAA<-649738+323313+48057+31053+20082+36227+29752+1443 -SPADs
tot<-SPADs+RAA
values<-c(rep(1,SPADs), rep(0,RAA))
# example values in subset
SPADs_subset<-5
n_subset<-36227
# Prepare for 100000 runs
reps<-100000
result<-logical(length=reps)
# Check side: greater or lesser than; if greater than expected:
if(SPADs_subset>(SPADs/tot*n_subset)){
  for(i in 1:reps){
    # draw random sample of subset size without replacement and check if as many or more values in drawn as in measured
  result[i]<-sum(sample(values, n_subset, replace = F)==1)>=SPADs_subset
  }
}
# Check side: greater or lesser than: if lesser than expected:
if(SPADs_subset<(SPADs/tot*n_subset)){
  for(i in 1:reps){
    # draw random sample of subset size without replacement and check if as little or less values in drawn as in measured
    result[i]<-sum(sample(values, n_subset, replace = F)==1)<=SPADs_subset
  }
}
p<-sum(result)/reps
p
```

Appendix C

	Behavioral data study	Accident data study
Period	20-08-2018 to 20-03-2020	01-01-2014 to 31-12-2019
Trains	Passenger trains from NS with the Orbit installed and active	All passenger trains, excluding those with Orbit
Train types	Regional trains underrepresented because one type of regional trains did not have Orbit installed yet	All passenger trains
Trains used in frequency calculation	All trains of the same train series mentioned at "Trains"	All trains of the same train series mentioned at "Trains"
Frequency calculation	All planned yellow+number aspects in the location for that train + yellow+number speed restrictions if the same yellow+number as planned yellow+number aspects	Frequency of one specific yellow+number aspect. Approaches are excluded when other yellow+number aspects or planned yellow are present more often than 5 times and/or the most prevalent yellow+number aspect is present less than 100 times in the previous 14 days.
Planned yellow-yellow-red approaches at location	Yellow-red approaches with planned yellow-yellow-red in the past fourteen days included, but yellow-yellow-red not used in the frequency calculated	Yellow-red approach filtered out if there is often a yellow as part of planned yellow-yellow-red in that location as per criteria at "Frequency calculation"

Table 11. Differences in data used between the study using behavioral data and the study using accident data

Appendix D

The signal chosen for this explorative analysis was the signal with the most red aspect approaches in the study described in chapter 2. A random date was chosen within the time period of data used in that study, namely 15-01-2019. We identified two train series that had more than two red aspect approaches towards that signal on that day. Next, for both train series, we counted the number of yellow aspects during a time periods of 14 days, specifically the first 14 days of the month and the last 14 days (day 01 to 14 and day 15 to 28). Data was gathered for January, March, May and July and thus 8 periods of 14 days.

Y:8	264	282	302	362	349	453	471	471
Y:4	2	4	1	4	3	3	4	4
Y	34	134	125	167	180	70	60	56
Time period in 2019	March 15-28	July 15-28	May 15-28	May 1-14	July 1-14	January 15-28	January 1-14	March 1-14

Table 12. No clear correlation between yellow+8 and yellow frequency for train series A.

Y:8	330	336	367	477	492	496	498	505
Y:4	2	5	7	5	2	4	5	5
Y	92	158	44	48	29	27	29	26
Time period in 2019	March 15-28	May 15-28	July 15-28	July 1-14	January 15-28	January 1-14	March 1-14	May 1-14

Table 13. No clear correlation between yellow+8 and yellow frequency for train series B.

Since there was no differentiation in the data analysis whether a yellow aspect was part of a yellow-yellow-red approach or not, it was checked whether there could actually be a scheduled yellow-yellow-red approach in this location. This is not the case. For the signal used to analyze the red aspect approaches, if the previous signal is 'yellow', then the next signal must either be red or yellow-flashing where yellow-flashing indicates that speed needs to be restricted to below 40 km/h and the driver should be able to stop at any point behind the signal, for example due to occupied tracks or danger. During January 01 to 28 of 2019, the aspect was never yellow-flashing at the location of the signal which was used for red aspect approach analysis. This indicates that there is never a scheduled yellow-yellow-red approach and there does not often seem to be a scheduled yellow-yellow+flashing approach. Thus, the presence of a yellow signal was usually followed by a red aspect, of which the moment of signal aspect improvement is unknown.

Overall, above data does not provide support for the hypothesis that higher yellow+number frequencies in the previous fourteen days are accompanied by a higher exposure to having to stop in front of the red aspect.

Appendix E

Suggested reading with respect to SPAD causation:

Gibson, H. (2016). Industry Human Factors SPAD Review – Project Summary Report. RSSB Human Factors. Retrievable from https://www.sparkrail.org/Lists/Records/DispForm.aspx?ID=22779

Turner, C., Harrison, R. & Lowe, E. (2003). Development of a human factors SPAD hazard checklist. *Contemporary Ergonomics 2003,* 385-390.

Turner, C. (2002). Human factors SPAD hazard checklist: Management summary. SPARK, retrievable from https://www.sparkrail.org/Lists/Records/DispForm.aspx?ID=20074

Hamilton, I. W. & Clarke, T. (2005). Driver performance modelling and its practical application to railway safety. *Applied Ergonomics, 36* (6), 661-670. doi: 10.1016/j.apergo.2005.07.005

Naweed, A., Rainbird, S. & Chapman, J. (2015). Investigating the formal countermeasures and informal strategies used to mitigate SPAD risk in train driving. Ergonomics, 58(6), Pages 883-896, doi:10.1080/00140139.2014.1001448

Naweed, A., Bowditch, L., Chapman, J. and Balfe, N. (2019). System precursors to signals passed at danger (SPADs): An exploratory comparison of SPAD history and rail environment. Conference paper at 12th World Congress on Railway Research (WCRR2019), Tokyo, Japan. Available at: https://www.researchgate.net/publication/337336060_System_precursors_to_signals_passed_at_danger_SPADs_An_exploratory_comparison_of_SPAD_history_and_rail_environment

Anjum Naweed (2020). Getting mixed signals: Connotations of teamwork as performance shaping factors in network controller and rail driver relationship dynamics. Applied Ergonomics, 82, doi:10.1016/j.apergo.2019.102976.

van den Top, J. (2010). Modelling Risk Control Measures in Railways: Analysing How Designers and Operators Organise Safe Rail Traffic. In; Next Generation Infrastructures Foundation: Delft, the Netherlands, ISBN 9789079787159.

Verstappen, V.J. (2017). The performance of Dutch train drivers based on the impact of the presence of a second person in the cab. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*., 231(10),1130-1140. doi:10.1177/0954409717694562

Suggested reading with respect to big data research and SPADs:

Rawia Ahmed Hassan E.L. Rashidy, Peter Hughes, Miguel Figueres-Esteban, Chris Harrison, Coen Van Gulijk, A big data modeling approach with graph databases for SPAD risk, Safety Science, Volume 110, Part B, 2018, Pages 75-79, ISSN 0925-7535, https://doi.org/10.1016/j.ssci.2017.11.019.

Figueres-Esteban, M., Hughes, P., El Rashidy, R. A. H. and Van Gulijk, C. (2017) Integrating data to support SPAD management. In: Sixth International Human Factors Rail Conference, 6-9 November 2017, London. (Unpublished) Available at http://eprints.hud.ac.uk/id/eprint/33962/

Zhao, Y., Stow, J. and Harrison, C. (2016). Improving the understanding of SPAD risks using red aspect approach data. Safety and Reliability, 36(3), p. 199-212, doi:10.1080/09617353.2016.1252086

Harrison, C., Stow, J., Ge, X., Gregory, J., Gibson, H. and Monk, A. (2022). At the limit? Using operational data to estimate train driver human reliability. Applied Ergonomics, Volume 104, doi:10.1016/j.apergo.2022.103795.

Zhao, Y., Stow, J. & Harrison, C. (2016). Improving the understanding of SPAD risks using red aspect approach data. *Safety and Reliabilty*, 36(3), 199-212. doi:10.1080/09617353.2016.1252086

Suggested reading with respect to big data research in railway safety:

Van Gulijk, C., Hughes, P., Figueres-Esteban, M., Dacre, M. and Harrison, C. (2015). Big Data Risk Analysis for Rail Safety? In: Safety and Reliability of Complex Engineered Systems: ESREL 2015. CRC/Balkema. ISBN 9781138028791

Parkinson, H.J. & Bamford, G. (2016). The Potential for Using Big Data Analytics to Predict Safety Risks by Analysing Rail Accidents. Proceedings of the Third International Conference on Railway Technology: Research, Development and Maintenance, J. Pombo, (Editor), Civil-Comp Press, Stirlingshire, Scotland. [2016]; paper 66;

D'Agostino, A. (2016). Big data in railways. European Union Agency for Railways, Safety Union. Retrievable from http://www.era.europa.eu/Document-Register/Pages/Big-data-in-railways.aspx

References

1. Concha-Barrientos, M.; Nelson, D.I.; Fingerhut, M.; Driscoll, T.; Leigh, J. The Global Burden Due to Occupational Injury. *Am J Ind Med* **2005**, *48*, 470–481, doi:10.1002/ajim.20226.
2. Eurostat Statistical Analysis of Socio-Economic Costs of Accidents at Work in the European Union. Working Papers and Studies. 2004.
3. Eurostat Work and Health in the EU. A Statistical Portrait. Panorama of the European Union. 2004.
4. Swuste, P.; van Gulijk, C.; Zwaard, W.; Lemkowitz, S.; Oostendorp, Y.; Groeneweg, J. *Van Veiligheid Naar Veiligheidskunde*; Vakmedianet: Alphen aan den Rijn, the Netherlands, 2019; ISBN 9789 4621 55817.
5. Reason, J. *Managing the Risks of Organizational Accidents*; Ashgate Publishing Limited: Surrey, England, 1997; ISBN 978 1 84014 104 7.
6. Wang, B.; Wang, Y. Big Data in Safety Management: An Overview. *Saf Sci* **2021**, *143*, 105414, doi:10.1016/j.ssci.2021.105414.
7. Ouyang, Q.; Wu, C.; Huang, L. Methodologies, Principles and Prospects of Applying Big Data in Safety Science Research. *Saf Sci* **2018**, *101*, 60–71, doi:10.1016/j.ssci.2017.08.012.
8. Inspectie Leefomgeving en Transport *Veiligheid van de Spoorwegen: Jaarverslag Spoorwegveiligheid 2019*; 2020;
9. Kyriakidis, M.; Simanjuntak, S.; Singh, S.; Majumdar, A. The Indirect Costs Assessment of Railway Incidents and Their Relationship to Human Error - The Case of Signals Passed at Danger. *Journal of Rail Transport Planning & Management* **2019**, *9*, 34–45, doi:10.1016/j.jrtpm.2019.01.001.
10. Naweed, A.; Trigg, J.; Cloete, S.; Allan, P.; Bentley, T. Throwing Good Money after SPAD? Exploring the Cost of Signal Passed at Danger (SPAD) Incidents to Australasian Rail Organisations. *Saf Sci* **2018**, *109*, 157–164, doi:10.1016/j.ssci.2018.05.018.
11. Ministry of Infrastructure and the Environment *Railway Map ERTMS Version 2.0 - State of Play Regarding Research in the Exploratory Phase*; 2013;
12. Smith, P.; Majumdar, A.; Ochieng, W.Y. An Overview of Lessons Learnt from ERTMS Implementation in European Railways. *Journal of Rail Transport Planning & Management* **2012**, *2*, 79–87, doi:10.1016/j.jrtpm.2013.10.004.
13. Ministerie van Infrastructuur en Waterstaat *Basisrapportage, Tevens Dertiende Voortgangsrapportage European Rail Traffic Management System (ERTMS)*; 2020;
14. BNR Webredactie CEO ProRail: 30 Procent Groei in 2030 2020.
15. GWW Grotere Capaciteit Spoorgoederenvervoer in 2030 En Nog Duurzaam Ook! Available online: gww-bouw.nl/artikel/grotere-capaciteit-spoorgoederenvervoer-in-2030-en-nog-duurzaam-ook (accessed on 28 April 2021).
16. Restrepo Ramos, F.D. Incidental Vocabulary Learning in Second Language Acquisition: A Literature Review. *PROFILE Issues in Teachers' Professional Development* **2015**, *17*, 157–166, doi:10.15446/profile.v17n1.43957.

17. Wagnon, C.C.; Wehrmann, K.; Klöppel, S.; Peter, J. Incidental Learning: A Systematic Review of Its Effect on Episodic Memory Performance in Older Age. *Front Aging Neurosci* **2019**, *10*, doi:10.3389/fnagi.2019.00173.

18. Macis, M.; Sonbul, S.; Alharbi, R. The Effect of Spacing on Incidental and Deliberate Learning of L2 Collocations. *System* **2021**, *103*, 102649, doi:10.1016/j.system.2021.102649.

19. Chen, Y. Comparing Incidental Vocabulary Learning from Reading-Only and Reading-While-Listening. *System* **2021**, *97*, 102442, doi:10.1016/j.system.2020.102442.

20. Webb, S.; Newton, J.; Chang, A. Incidental Learning of Collocation. *Lang Learn* **2013**, *63*, 91–120, doi:10.1111/j.1467-9922.2012.00729.x.

21. Hulstijn, J.H. Incidental Learning in Second Language Acquisition. In *The Encyclopedia of Applied Linguistics*; Blackwell Publishing Ltd: Oxford, UK, 2012.

22. Eysenck, M.W. Age Differences in Incidental Learning. *Dev Psychol* **1974**, *10*, 936–941, doi:10.1037/h0037263.

23. Paller, K.A.; Kutas, M.; Mayes, A.R. Neural Correlates of Encoding in an Incidental Learning Paradigm. *Electroencephalogr Clin Neurophysiol* **1987**, *67*, 360–371, doi:10.1016/0013-4694(87)90124-6.

24. Heinström, J. Psychological Factors behind Incidental Information Acquisition. *Libr Inf Sci Res* **2006**, *28*, 579–594, doi:10.1016/j.lisr.2006.03.022.

25. Greene, J.A.; Copeland, D.Z.; Deekens, V.M. A Model of Technology Incidental Learning Effects. *Educ Psychol Rev* **2021**, *33*, 883–913, doi:10.1007/s10648-020-09575-5.

26. Marsick, V.J.; Watkins, K.E.; Scully-Russ, E.; Nicolaides, A. Rethinking Informal and Incidental Learning in Terms of Complexity and the Social Context. *Journal of Adult Learning, Knowledge and Innovation* **2017**, *1*, 27–34, doi:10.1556/2059.01.2016.003.

27. Gibson, H.; Roels, R.; Harrison, C.; Kohli, R. Underlying Causes of Signals Passed at Danger - Looking at a Spike. In Proceedings of the 7th International Human Factors Virtual Rail Conference 2021; 2021.

28. Baysari, M.T.; McIntosh, A.S.; Wilson, J.R. Understanding the Human Factors Contribution to Railway Accidents and Incidents in Australia. *Accid Anal Prev* **2008**, *40*, 1750–1757, doi:10.1016/j.aap.2008.06.013.

29. Madigan, R.; Golightly, D.; Madders, R. Application of Human Factors Analysis and Classification System (HFACS) to UK Rail Safety of the Line Incidents. *Accid Anal Prev* **2016**, *97*, 122–131, doi:10.1016/j.aap.2016.08.023.

30. Wickens, C.D.; Lee, J.; Liu, Y.D.; Gordon-Becker, S. *Introduction to Human Factors Engineering*; 2nd ed.; Pearson Education (US): Upper Saddle River, NJ, USA, 2003;

31. Wilson, J.R.; Norris, B.J. Rail Human Factors: Past, Present and Future. *Appl Ergon* **2005**, *36*, 649–660, doi:10.1016/j.apergo.2005.07.001.

32. Hamilton, W.I.; Clarke, T. Driver Performance Modelling and Its Practical Application to Railway Safety. *Appl Ergon* **2005**, *36*, 661–670, doi:10.1016/j.apergo.2005.07.005.

33. Lawton, R.; Ward, N.J. A Systems Analysis of the Ladbroke Grove Rail Crash. *Accid Anal Prev* **2005**, *37*, 235–244, doi:10.1016/j.aap.2004.08.001.

34. Naweed, A.; Bowditch, L.; Chapman, J.; Balfe, N.; Dorrian, J. System Precursors to Signals Passed at Danger (SPADs): An Exploratory Comparison of SPAD History and Rail Environment. **2019**, 1–9.

35. Stanton, N.A.; Walker, G.H. Exploring the Psychological Factors Involved in the Ladbroke Grove Rail Accident. *Accid Anal Prev* **2011**, *43*, 1117–1127, doi:10.1016/j.aap.2010.12.020.

36. Punzet, L.; Pignata, S.; Rose, J. Error Types and Potential Mitigation Strategies in Signal Passed at Danger (SPAD) Events in an Australian Rail Organisation. *Saf Sci* **2018**, *110*, 89–99, doi:10.1016/j.ssci.2018.05.015.

37. Naweed, A.; Hockey, G.R.J.; Clarke, S.D. Designing Simulator Tools for Rail Research: The Case Study of a Train Driving Microworld. *Appl Ergon* **2013**, *44*, 445–454, doi:10.1016/j.apergo.2012.10.005.

38. van den Top, J. Modelling Risk Control Measures in Railways: Analysing How Designers and Operators Organise Safe Rail Traffic. In; Next Generation Infrastructures Foundation: Delft, the Netherlands, 2010 ISBN 9789079787159.

39. Hebb, D.O. *The Organization of Behavior*; Wiley & Sons: New York, NY, USA, 1949;

40. Gazzaniga, M.; Ivry, R.; Mangun, G. Memory and Brain. In *Cognitive Neuroscience: The Biology of the Mind*; W.W. Norton & Company, 2009; pp. 324–356.

41. Broadbent, D.E. Effective Decisions and Their Verbal Justification. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* **1990**, *327*, 493–502, doi:10.1098/rstb.1990.0092.

42. Dienes, D.; Perner, J. A Theory of the Implicit Nature of Implicit Learning. In *Implicit Learning and Consciousness*; French, R.M., Cleeremans, A., Eds.; Psychology Press: East Sussex, UK, 2002; pp. 68–92 ISBN 978-1-138-87741-2.

43. Ebbinghaus, H. *Über Das Gedächtnis. Untersuchungen Zur Experimentellen Psychologie*; Scientia Verlag, 1885;

44. Wang, B.; Theeuwes, J. Implicit Attentional Biases in a Changing Environment. *Acta Psychol (Amst)* **2020**, *206*, 103064, doi:10.1016/j.actpsy.2020.103064.

45. Bartlett, F. *Remembering: A Study in Experimental and Social Psychology*; Cambridge University Press: Cambridge, UK, 1932;

46. Norman, D.A. Categorization of Action Slips. *Psychol Rev* **1981**, *88*, 1–15, doi:10.1037/0033-295X.88.1.1.

47. Piaget, J.; Cook, M.T. *The Origins of Intelligence in Children*; International University Press: New York, NY, USA, 1952;

48. Wrisberg, C.A.; Shea, C.H. Shifts in Attention Demands and Motor Program Utilization During Motor Learning. *J Mot Behav* **1978**, *10*, 149–158, doi:10.1080/00222895.1978.10735148.

49. Näätänen, R. Brain Physiology and the Unconscious Initiation of Movements. *Behavioral and Brain Sciences* **1985**, *8*, 549–549, doi:10.1017/S0140525X00045039.

50. D'Ostilio, K.; Garraux, G. Brain Mechanisms Underlying Automatic and Unconscious Control of Motor Action. *Front Hum Neurosci* **2012**, *6*, doi:10.3389/fnhum.2012.00265.

51. Pessiglione, M.; Schmidt, L.; Draganski, B.; Kalisch, R.; Lau, H.; Dolan, R.J.; Frith, C.D. How the Brain Translates Money into Force: A Neuroimaging Study of Subliminal Motivation. *Science (1979)* **2007**, *316*, 904–906, doi:10.1126/science.1140459.

52. Schmidt, L.; Palminteri, S.; Lafargue, G.; Pessiglione, M. Splitting Motivation: Unilateral Effects of Subliminal Incentives. *Psychol Sci* **2010**, *21*, 977–983, doi:10.1177/0956797610372636.

53. Biran, I.; Giovannetti, T.; Buxbaum, L.; Chatterjee, A. The Alien Hand Syndrome: What Makes the Alien Hand Alien? *Cogn Neuropsychol* **2006**, *23*, 563–582, doi:10.1080/02643290500180282.

54. Weiskrantz, L.; Warrington, E.K.; Sanders, M.D.; Marshall, J. Visual Capacity in the Hemianopic Field Following a Restricted Occipital Ablation. *Brain* **1974**, *97*, 709–728, doi:10.1093/brain/97.1.709.

55. Scarpina, F.; Tagini, S. The Stroop Color and Word Test. *Front Psychol* **2017**, *8*, doi:10.3389/fpsyg.2017.00557.

56. Neumann, O. Automatic Processing: A Review of Recent Findings and a Plea for an Old Theory. In *Cognition and Motor Processes*; Prinz, W., Sanders, A.F., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, Germany, 1984; pp. 255–293.

57. Wiegmann, D.A.; Shappell, S.A. *A Human Error Approach to Aviation Accident Analysis: The Human Factors Analysis and Classification System.*; CRC Press LLC: Abingdon, UK, 2003;

58. Gertman, D.I.; Blackman, H.S.; Marble, J.L.; Smith, C.; Boring, R.L.; O'Reilly, P. The SPAR H Human Reliability Analysis Method. *American Nuclear Society 4th International Topical Meeting on Nuclear Plant Instrumentation, Control and Human Machine Interface Technology* **2004**, 17–24.

59. Lowe, T.; Turner, C. A Human Factors SPAD Checklist. In Proceedings of the Proceedings of the First European Conference on Rail Human Factors; York, UK, 2003.

60. Independent Transport Safety Regulator *Mitigation Measures for Tool C - Rail Infrastructure Managers*; 2013;

61. Gibson, H. *Industry Human Factors SPAD Review - Project Summary Report*; 2016;

62. Naweed, A.; Aitken, J. Drive a Mile in My Seat: Signal Design from a Systems Perspective. *Technical Meeting of the Institution of Railway Signal Engineers* **2014**, 1–7.

63. Balfe, N.; Geoghegan, S.; Smith, B. SPAD Dashboard: A Tool for Tracking and Analysing Factors Influencing SPADs. In Proceedings of the Contemporary Ergonomics and Human Factors; Charles, R., Wilkinson, J., Eds.; CIEHF, 2017.

64. Singer, M.H.; Lappin, J.S. Similarity: Its Definition and Effect on the Visual Analysis of Complex Displays. *Percept Psychophys* **1976**, *19*, 405–411, doi:10.3758/BF03199400.

65. NS Werft Jaarlijks 250 Machinisten Available online: https://www.treinreiziger.nl/ns-werft-jaarlijks-250-machinisten/ (accessed on 3 December 2020).

66. Standplaats Available online: https://wiki.ovinnederland.nl/wiki/Standplaats (accessed on 3 December 2020).

67. Jacobs, I. *OVPRO*. 2016,.

68. Slidestops Hoeveel Trajecten Rijdt Een Machinist Eigenlijk? Available online: https://community.ns.nl/off-topic-32/hoeveel-trajecten-rijdt-een-machinist-eigenlijk-36445?postid=235209#post235209 (accessed on 2 July 2021).

69. Burggraaf, J.; Groeneweg, J.; Sillem, S.; van Gelder, P. What Employees Do Today Because of Their Experience Yesterday: How Incidental Learning Influences Train Driver Behavior and Safety Margins (A Big Data Analysis). *Safety* **2021**, *7*, 2, doi:10.3390/safety7010002.

70. Balfe, N.; Doyle, K. Delving Deeper: Applying Human Factors Analysis to Identify Factors Contributing to Railway Incidents. In Proceedings of the 7th International Human Factors Virtual Rail Conference 2021; Virtual, 2021.

71. Norman, D.A.; Shallice, T. Attention to Action. In *Consciousness and Self-Regulation*; Springer US: Boston, MA, 1986; pp. 1–18.

72. ProRail *Ontwerpvoorschrift: Plaatsing En Toepassing van Lichtseinen*; 2020;

73. Summerfield, C.; Egner, T. Expectation (and Attention) in Visual Cognition. *Trends Cogn Sci* **2009**, *13*, 403–409, doi:10.1016/j.tics.2009.06.003.

74. Gibson, W.H.; Willet, J.; Lewis, G.; Harrison, C. Exploring the Limits of Train Driver Reliability. In Proceedings of the Sixth International Human Factors Rail Conference; London, England, 2017.

75. Harrison, C.; Stow, J.; Ge, X.; Gregory, J.; Gibson, H. Using the Red Aspect Approaches To Signals (RAATS) Tool to Better Understand the Human Relaibility Associated with SPADs. In Proceedings of the 7th International Human Factors Virtual Rail Conference 2021; 2021.

76. Buttle, H.; Raymond, J.E. High Familiarity Enhances Visual Change Detection for Face Stimuli. *Percept Psychophys* **2003**, *65*, 1296–1306, doi:10.3758/BF03194853.

77. Ning, R. How Language Proficiency Influences Stroop Effect and Reverse-Stroop Effect: A Functional Magnetic Resonance Imaging Study. *J Neurolinguistics* **2021**, *60*, 101027, doi:10.1016/j.jneuroling.2021.101027.

78. Spering, M.; Carrasco, M. Acting without Seeing: Eye Movements Reveal Visual Processing without Awareness. *Trends Neurosci* **2015**, *38*, 247–258, doi:10.1016/j.tins.2015.02.002.

79. Heath, M.; Maraj, A.; Godbolt, B.; Binsted, G. Action Without Awareness: Reaching to an Object You Do Not Remember Seeing. *PLoS One* **2008**, *3*, e3539, doi:10.1371/journal.pone.0003539.

80. Chan, D.; Peterson, M.A.; Barense, M.D.; Pratt, J. How Action Influences Object Perception. *Front Psychol* **2013**, *4*, doi:10.3389/fpsyg.2013.00462.

81. Kiefer, M.; Trumpp, N.M. Embodiment Theory and Education: The Foundations of Cognition in Perception and Action. *Trends Neurosci Educ* **2012**, *1*, 15–20, doi:10.1016/j.tine.2012.07.002.

82. Vernon, D.; Lowe, R.; Thill, S.; Ziemke, T. Embodied Cognition and Circular Causality: On the Role of Constitutive Autonomy in the Reciprocal Coupling of Perception and Action. *Front Psychol* **2015**, *6*, doi:10.3389/fpsyg.2015.01660.

83. Martens, M.H. The Failure to Respond to Changes in the Road Environment: Does Road Familiarity Play a Role? *Transp Res Part F Traffic Psychol Behav* **2018**, *57*, 23–35, doi:10.1016/j.trf.2017.08.003.

84. van Schaardenburg-Verhoeve, K. Analyse van Ongevallen: Tradities, Vernieuwing En Toepassing van Modellen En Methoden. In *Methodische aspecten van het onderzoek naar ongevallen*; Mertens, F., van Schaardenburg-Verhoeve, K., Sillem, S., Eds.; Eburon, Delft, 2012; pp. 15–32 ISBN 978-90-5972-700-7.

85. Salmon, P.M.; Cornelissen, M.; Trotter, M.J. Systems-Based Accident Analysis Methods: A Comparison of Accimap, HFACS, and STAMP. *Saf Sci* **2012**, *50*, 1158–1170, doi:10.1016/j.ssci.2011.11.009.

86. Burggraaf, J.; Groeneweg, J. Managing the Human Factor in the Incident Investigation Process. In Proceedings of the SPE International Conference and Exhibition on Health, Safety, Security, Environment, and Social Responsibility; Society of Petroleum Engineers, April 11 2016.

87. Dutch Safety Board *Misaligned Take-off from Runway 24, Amsterdam Airport Schiphol*; The Hague, 2018;

88. Dutch Safety Board *Stuwaanvaring Door Benzeentanker Bij Grave*; The Hague, the Netherlands, 2018;

89. National Quality Board *Human Factors in Healthcase: A Concordat from the National Quality Board*;

90. Bäckman, L.; Nilsson, L.-G.; Chalom, D. New Evidence on the Nature of the Encoding of Action Events. *Mem Cognit* **1986**, *14*, 339–346, doi:10.3758/BF03202512.

91. Craik, F.I.M.; Govoni, R.; Naveh-Benjamin, M.; Anderson, N.D. The Effects of Divided Attention on Encoding and Retrieval Processes in Human Memory. *J Exp Psychol Gen* **1996**, *125*, 159–180, doi:10.1037/0096-3445.125.2.159.

92. Monk, A. Mode Errors: A User-Centred Analysis and Some Preventative Measures Using Keying-Contingent Sound. *Int J Man Mach Stud* **1986**, *24*, 313–327, doi:10.1016/S0020-7373(86)80049-9.

93. Bernstein, P.L. *Against the Gods. The Remarkable Story of Risk*; John Wiley & Sons Inc.: New York, 1998;

94. Shewhart, W.A. *Economic Control of Quality of Manufactured Product*; seventh.; D. van Nostrand Company, Inc., 1931;

95. Montgomery, D.C. Chapter 5 Methods and Philosophy of Statistical Process Control. In *Statistical Quality Control*; John Wiley & Sons Inc., 2012.

96. Hopkins, A. Thinking About Process Safety Indicators. *Saf Sci* **2009**, *47*, 460–465, doi:10.1016/j.ssci.2007.12.006.

97. Øien, K.; Utne, I.B.; Herrera, I.A. Building Safety Indicators: Part 1 – Theoretical Foundation. *Saf Sci* **2011**, *49*, 148–161, doi:10.1016/j.ssci.2010.05.012.

98. Best, M. Walter A Shewhart, 1924, and the Hawthorne Factory. *Qual Saf Health Care* **2006**, *15*, 142–143, doi:10.1136/qshc.2006.018093.

99. Perez, P.; Holloway, J.; Ehrenfeld, L.; Cohen, S.; Cunningham, L.; Miley, G.B.; Hollenbeck, B.L. Door Openings in the Operating Room Are Associated with Increased Environmental Contamination. *Am J Infect Control* **2018**, *46*, 954–956, doi:10.1016/j.ajic.2018.03.005.

100. Roth, J.A.; Juchler, F.; Dangel, M.; Eckstein, F.S.; Battegay, M.; Widmer, A.F. Frequent Door Openings During Cardiac Surgery Are Associated With Increased Risk for Surgical Site Infection: A Prospective Observational Study. *Clinical Infectious Diseases* **2019**, *69*, 290–294, doi:10.1093/cid/ciy879.

101. Young, R.S.; O'Regan, D.J. Cardiac Surgical Theatre Traffic: Time for Traffic Calming Measures? *Interact Cardiovasc Thorac Surg* **2010**, *10*, 526–529, doi:10.1510/icvts.2009.227116.

102. Prakken, F.J.; Lelieveld-Vroom, G.M.M.; Milinovic, G.; Jacobi, C.E.; Visser, M.J.T.; Steenvoorde, P. Meetbaar Verband Tussen Preventieve Interventies En de Incidentie van Postoperatieve Wondinfecties. *Nederlands Tijdschrift Geneeskunde* **2011**, *155*, 1–6.

103. Fisher, C.D. What If We Took Within-Person Performance Variability Seriously? *Ind Organ Psychol* **2008**, *1*, 185–189, doi:10.1111/j.1754-9434.2008.00036.x.

104. Mears, S.C.; Blanding, R.; Belkoff, S.M. Door Opening Affects Operating Room Pressure During Joint Arthroplasty. *Orthopedics* **2015**, *38*, e991–e994, doi:10.3928/01477447-20151020-07.

105. Reid, D.E. Determining the Optimal Turnout Time Standards for the Stanwood Fire Department. **2011**.

106. National Fire Protection Association, . *NFPA 1710: Standard for Organization and Deployment of Fire Suppression Operations, Emergency Medical Operations, and Special Operations for the Public by Career Fire Departments*; 2010;

107. Upson, R.; Notarianni, K.A. *Quantitative Evaluation of Fire and EMS Mobilization Times*; Springer, 2010;

108. Reglen, D.; Scheller, D.S. Fire Department Turnout Times: A Contextual Analysis. *J Homel Secur Emerg Manag* **2016**, *13*, doi:10.1515/jhsem-2015-0015.

109. Sammer, C.E.; Lykens, K.; Singh, K.P.; Mains, D.A.; Lackan, N.A. What Is Patient Safety Culture? A Review of the Literature. *Journal of Nursing Scholarship* **2010**, *42*, 156–165, doi:10.1111/j.1547-5069.2009.01330.x.

110. Ziemann, M.; Eren, Y.; El-Osta, A. Gene Name Errors Are Widespread in the Scientific Literature. *Genome Biol* **2016**, *17*, 177, doi:10.1186/s13059-016-1044-7.

111. Eklund, A.; Nichols, T.E.; Knutsson, H. Cluster Failure: Why FMRI Inferences for Spatial Extent Have Inflated False-Positive Rates. *Proceedings of the National Academy of Sciences* **2016**, *113*, 7900–7905, doi:10.1073/pnas.1602413113.

112. Bird, J. Bugs and Numbers: How Many Bugs Do You Have in Your Code? Available online: http://swreflections.blogspot.nl/2011/08/bugs-and-numbers-how-many-bugs-do-you.html.

113. Garfunkel, S. History's Worst Software Bugs Available online: https://archive.wired.com/software/coolapps/news/2005/11/69355?currentPage=all.

114. Kaplan, R.M.; Chambers, D.A.; Glasgow, R.E. Big Data and Large Sample Size: A Cautionary Note on the Potential for Bias. *Clin Transl Sci* **2014**, *7*, 342–346, doi:10.1111/cts.12178.

115. Cai, L.; Zhu, Y. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Sci J* **2015**, *14*, 2, doi:10.5334/dsj-2015-002.

116. Lovelace, R.; Birkin, M.; Cross, P.; Clarke, M. From Big Noise to Big Data: Toward the Verification of Large Data Sets for Understanding Regional Retail Flows. *Geogr Anal* **2016**, *48*, 59–81, doi:10.1111/gean.12081.

117. Otero, C.E.; Peter, A. Research Directions for Engineering Big Data Analytics Software. *IEEE Intell Syst* **2015**, *30*, 13–19, doi:10.1109/MIS.2014.76.

118. Morewedge, C.K.; Kahneman, D. Associative Processes in Intuitive Judgment. *Trends Cogn Sci* **2010**, *14*, 435–440, doi:10.1016/j.tics.2010.07.004.

119. Kahneman, D. *Thinking, Fast and Slow*; Penguin Books Ltd: London, England, 2011;

120. Baybutt, P. Cognitive Biases in Process Hazard Analysis. *J Loss Prev Process Ind* **2016**, *43*, 372–377, doi:10.1016/j.jlp.2016.06.014.

121. Tversky, A.; Kahneman, D. Judgment under Uncertainty: Heuristics and Biases. *Science (1979)* **1974**, *185*, 1124–1131, doi:10.1126/science.185.4157.1124.

122. Trimmer, P.C. Optimistic and Realistic Perspectives on Cognitive Biases. *Curr Opin Behav Sci* **2016**, *12*, 37–43, doi:10.1016/j.cobeha.2016.09.004.

123. Mohanani, R.; Salman, I.; Turhan, B.; Rodriguez, P.; Ralph, P. Cognitive Biases in Software Engineering: A Systematic Mapping Study. **2017**, doi:10.1109/TSE.2018.2877759.

124. Haselton, M.G.; Bryant, G.A.; Wilke, A.; Frederick, D.A.; Galperin, A.; Frankenhuis, W.E.; Moore, T. Adaptive Rationality: An Evolutionary Perspective on Cognitive Bias. *Soc Cogn* **2009**, *27*, 733–763, doi:10.1521/soco.2009.27.5.733.

125. Blumenthal-Barby, J.S.; Krieger, H. Cognitive Biases and Heuristics in Medical Decision Making. *Medical Decision Making* **2015**, *35*, 539–557, doi:10.1177/0272989X14547740.

126. Clarke, D.D.; Sokoloff, L. The Brain Consumes about One-Fifth of Total Body Oxygen. In *Basic Neurochemistry: Molecular, Cellular and Medical Aspects*; Siegel, G.W., Agranoff, W.B., Albers, R.W., Eds.; Lippincott-Raven, 1999.

127. Kuzawa, C.W.; Chugani, H.T.; Grossman, L.I.; Lipovich, L.; Muzik, O.; Hof, P.R.; Wildman, D.E.; Sherwood, C.C.; Leonard, W.R.; Lange, N. Metabolic Costs and Evolutionary Implications of Human Brain Development. *Proceedings of the National Academy of Sciences* **2014**, *111*, 13010–13015, doi:10.1073/pnas.1323099111.

128. Pronin, E.; Lin, D.Y.; Ross, L. The Bias Blind Spot: Perceptions of Bias in Self Versus Others. *Pers Soc Psychol Bull* **2002**, *28*, 369–381, doi:10.1177/0146167202286008.

129. Pronin, E. Perception and Misperception of Bias in Human Judgment. *Trends Cogn Sci* **2007**, *11*, 37–43, doi:10.1016/j.tics.2006.11.001.

130. Haugen, N.C. An Empirical Study of Using Planning Poker for User Story Estimation. In Proceedings of the AGILE 2006 (AGILE'06); IEEE; pp. 23–34.

131. Stanovich, K.E.; West, R.F. On the Relative Independence of Thinking Biases and Cognitive Ability. *J Pers Soc Psychol* **2008**, *94*, 672–695, doi:10.1037/0022-3514.94.4.672.

132. Neely, J.H. Semantic Priming and Retrieval from Lexical Memory: Roles of Inhibitionless Spreading Activation and Limited-Capacity Attention. *J Exp Psychol Gen* **1977**, *106*, 226–254, doi:10.1037/0096-3445.106.3.226.

133. Oswald, M.E.; Grosjean, S. Confirmation Bias. In *Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement and Memory*; Pohl, R.F., Ed.; Psychology Press: Hove, UK, 2004; pp. 79–96 ISBN 978-1-84169-351-4, OCLC 55124398.

134. Olson, E.A. "You Don't Expect Me to Believe That, Do You?" Expectations Influence Recall and Belief of Alibi Information. *J Appl Soc Psychol* **2013**, *43*, 1238–1247, doi:10.1111/jasp.12086.

135. Dougherty, M.R.P.; Gettys, C.F.; Ogden, E.E. MINERVA-DM: A Memory Processes Model for Judgments of Likelihood. *Psychol Rev* **1999**, *106*, 180–209, doi:10.1037/0033-295X.106.1.180.

136. Hernandez, I.; Preston, J.L. Disfluency Disrupts the Confirmation Bias. *J Exp Soc Psychol* **2013**, *49*, 178–182, doi:10.1016/j.jesp.2012.08.010.

137. Yin, R.K. *Case Study Research Design and Methods: Applied Social Research and Methods Series.*; second.; Sage Publications, Inc: Thousand Oaks, CA, 1994;

138. Leary, M.R. *Introduction to Behavioral Research Methods.*; Fifth.; Pearson Education (US), 2008;

139. van Gelder, P.H.A.J.M.; Vrijling, J.K. Homogeneity Aspects in Statistical Analysis of Coastal Engineering Data. *Coastal Engineering* **1998**, *26*, 3215–3223.

140. van Gelder, P.H.A.J.M.; Nijs, M. Statistical Flaws in Design and Analysis of Fertility Treatment Studies on Cryopreservation Raise Doubts on the Conclusions. *Facts Views Vis Obgyn* **2011**, *3*, 273–280.

141. Lazer, D.; Kennedy, R.; King, G.; Vespignani, A. The Parable of Google Flu: Traps in Big Data Analysis. *Science (1979)* **2014**, *343*, 1203–1205, doi:10.1126/science.1248506.

142. Figueres-Esteban, M.; Hughes, P.; van Gulijk, C. The Role of Data Visualization in Railway Big Data Risk Analysis. In *Safety and Reliability of Complex Engineered Systems*; CRC Press, 2015; pp. 2877–2882.