Visual Attention in Human−Machine Interaction

Eisma, Y.B.

DOI
10.4233/uuid:389a033a-88cc-433f-bbcb-1cd172c1ac0b

Publication date
2021

Document Version
Final published version

Citation (APA)
Eisma, Y. B. (2021). *Visual Attention in Human−Machine Interaction*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:389a033a-88cc-433f-bbcb-1cd172c1ac0b

VISUAL ATTENTION IN HUMAN–MACHINE INTERACTION

Yke Bauke Eisma

Visual Attention in Human–Machine Interaction

Yke Bauke Eisma

Layout and design: Eduard Boxem, persoonlijkproefschrift.nl

Printing: Gildeprint Enschede, gildeprint.nl

Visual Attention in Human−Machine Interaction

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 for Doctorates
to be defended publicly on
Friday 29 March 2021 at 15:00 o'clock

by

Yke Bauke EISMA

Master of Science in Mechanical Engineering, Delft University of Technology,
the Netherlands

Born in Wâlterswâld, the Netherlands

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus	chairperson
Dr. ir. J.C.F. de Winter	Delft University of Technology, promotor
Dr. ir. M.M. van Paassen	Delft University of Technology, promotor

Independent members:

Prof. dr. M. P. Hagenzieker	Delft University of Technology
Prof. dr. D. A. de Waard	University of Groningen
Prof. dr. ir. J. M. Hoekstra	Delft University of Technology
Prof. dr. I. Horvath	Delft University of Technology
Prof. dr. G. Matthews	University of Central Florida
Prof. dr. ir. D. A. Abbink	Delft University of Technology, reserve member

A large part of this thesis was performed within the research program VIDI with grant number TTW 016.Vidi.178.047 (2018–2022; “How should automated vehicles communicate with other road users?”), financed by the Netherlands Organisation for Scientific Research (NWO).

TABLE OF CONTENTS

VISUAL ATTENTION IN HUMAN–MACHINE INTERACTION

Summary

Humans are incapable of attending to everything at the same time. The serial nature of focused attention limits the information intake capacity of the perceptual system.

This thesis deals with the measurement and modelling of visual attention distribution. It is examined whether measures of visual attention are predictive of task performance.

Chapter 1: Introduction

The first chapter introduces the main topic of this thesis: the complex nature of modern technological systems, which feature many information sources that have to be monitored.

Many psychological constructs have been proposed in the human factors literature that have alleged criterion validity for task performance. Here, task performance is regarded as the human's ability to e.g., take over control of an automated system in potential critical situations. Contrary to the speculative nature of some of the Human-Factors constructs, this thesis sets out to capture performance in terms of objective measures of visual attention.

Wickens's (2008) Salience, Effort, Expectancy, Value (SEEV) model is introduced and discussed. This model is utilized for interpreting the eye-tracking results. Finally, a rationale for the topics in the thesis is provided. Chapters 2 through 4 of this thesis discuss and elaborate on Senders's (1983) research in detail, by means of replication research and an extensive tutorial on his mathematical models. These chapters provide an empirical underpinning and conceptual understanding of the concept of visual attention. Chapters 5 through 8 discuss visual attention in light of Air Traffic Control (ATC) and automated driving, and are regarded as suitable cases for attention distribution measurement and task performance prediction. Chapter 9 investigates task performance and visual attention in a psychometric task: Inspection Time, which provides a good testbed for operationalizing the effect of attention on task performance. Chapter 10 concludes with a discussion on the topics in this thesis.

Chapters 2-3: Visual attention in a dial-monitoring task

Chapters 2 and 3 elaborate on seminal experimental and theoretical work of John Senders (1983) on visual attention distribution in a dial-monitoring task. Participants had to detect threshold crossings of dial pointers that moved at different speeds (i.e., bandwidths).

Chapter 2 focusses on the experimental replication of Senders's work, whereas Chapter 3 elaborates on the quantitative mathematical models of visual sampling that Senders proposed. In Chapter 2, it was shown that Senders's original results from 1964 were replicated with high accuracy. Furthermore, in the replication experiment, it was shown

that visual sampling depends not only on the information bandwidth of the stimulus (Expectancy) but also on the Saliency of the stimulus and the Effort it requires to scan the entirety of the stimulus.

Chapter 3 interprets the results of Chapter 2 mathematically and explains and clarifies Senders's original mathematical models. Bandwidth-dependent (i.e., Expectancy-dependent) sampling is described through the Periodic Sampling Model and the Random Constrained Sampling Model. Contextual effects, such as Saliency and prior knowledge of the signal, are accounted for in the Conditional Sampling Model. Based on Chapters 2 and 3, it is concluded that Senders's work has good criterion validity for predicting visual sampling processes in simple monitoring tasks.

Chapter 4: Towards the use of visual attention for measuring situation awareness

In line with the aim of this thesis, that is, to find a measure that predicts task performance, Chapter 4 provides a critical narrative concerning a seminal construct in the Human Factors literature: Situation Awareness. Situation Awareness, a construct formalized and operationalized by Mica Endsley (1987, 1995), has reported criterion validity for task performance in a broad spectrum of application areas (e.g., ATC, flying, and car driving). Chapter 4 expands the discussion on the experiment of Chapter 2, in which we also administered a frequently used Situation Awareness measurement technique: SAGAT. It is concluded that SAGAT has modest criterion validity for task performance. Eye-movements were significantly more predictive of task performance than the SAGAT measure. The chapter concludes with a discussion on the pragmatic application of eye-tracking to measure Situation Awareness.

Chapters 5-6: Visual attention in air traffic control

Chapters 5 and 6 focus on task performance and visual attention distribution in the context of ATC.

Chapter 5 describes a study in which participants were subjected to an ATC-like conflict detection task, in which they had to continuously indicate whether two moving objects were on a collision course. Dependent variables were eye movements and spacebar pressing (i.e., conflict detection) performance. Independent variables were the conflict angle (30, 100, 150 degrees), update rate (continuous versus discrete), and conflict occurrence. Results indicated that 30-degree angles yielded the best performance and 100-degree the worst. Furthermore, discrete stimuli yielded a worse performance than continuous update rate stimuli. The higher performance on shallow conflict angles may be explained by perceptual heuristics, such as the 'closer is first' strategy. Eye-movement analysis confirmed this heuristic-based hypothesis for shallow angles, as participants employed smooth pursuit eye-movements, whereas for larger conflict angles participants mainly employed back-and-forth sampling between aircraft and

conflict point. Eye-movements patterns are thus for a large part explainable in terms of the distance between the dots, which is larger when the conflict angle is larger, a hypothesis that is in line with the SEEV model.

Chapter 6 investigated conflict detection performance, however, this time in a static ATC-like task. Therein, it was the goal to evaluate the effect of augmented feedback (a so-called Solution Space Diagram; SSD) on participants' workload, performance, and visual attention distribution. The results indicated that the augmented feedback condition resulted in lower self-reported task difficulty and a higher conflict detection rate. False-positive rates were approximately equal between groups. Furthermore, the SSD group participants spent a large portion of time looking at the SSD at the expense of looking at other task-relevant parts of the visual scene.

Chapters 7-8: Visual attention in the perception of automated vehicles

Chapters 7 and 8 aim to operationalize task performance and visual attention distribution in the interaction between pedestrians and automated vehicles (AVs). Both chapters focus on the use of external Human-Machine Interfaces (eHMIs) for AV–pedestrian communication.

Chapter 7 describes an experiment in which different eHMI placements were evaluated in terms of participant's spacebar pressing behavior (as an index of when participants felt safe to cross) and eye-movements. The independent variable was eHMI placement (roof, windscreen, grill, above the wheels, or a projection on the road). Results indicated that when the car slowed down, the roof, windscreen, and grill eHMIs yielded superior performance compared to the projection and wheels eHMIs. Eye-movement analysis revealed that the projection eHMI yielded more dispersed eye-movements than the other eHMIs, indicating that participants scanned more back and forth between eHMI and other relevant features of the scene. It was concluded that eHMIs should be mounted on different sides of the vehicle for optimal visibility.

Chapter 8 focusses on the so-called message perspective of the eHMI's message, in other words, whether an eHMI should feature an instructive message (i.e., 'WALK', or 'DON'T WALK', also: *egocentric* messages) or an informative message (i.e., 'DRIVING' or 'BRAKING', also: *allocentric* messages) for optimal communicational clarity. Also, the effect of ambiguous messages (i.e., 'STOP' and 'GO') was investigated in terms of eye-movements and response performance. Participants were asked to respond with 'yes' or 'no' (left and right shift keys) to the statement 'I can cross' when presented with a photo of a car that featured a car with an eHMI displaying one of the aforementioned eHMIs messages. A memory task was included to simulate the effect of real-life workload. The experiment results revealed that egocentric messages were most persuasive, demonstrated by more consistent crossing decisions and faster response times. Furthermore, the results indicated that the ambiguous messages were interpreted

from an egocentric perspective, that is, 'GO' encouraged crossing, and 'STOP' inhibited crossing. Eye-movement analyses revealed that longer text messages caused a higher number of saccades, but did not inhibit task performance. It is concluded that eHMIs may have to feature egocentric messages.

Chapter 9: Visual attention in a psychometrics task

Chapter 9 studies visual attention distribution in the context of an elementary psychometrics task: Inspection Time (IT). Here, the effect of different stimulus exposure times was investigated on task performance, as measured by response accuracy and response times. Furthermore, the effect of higher-order strategies and perception of visual illusions were evaluated in the context of task performance. Two large-sample experiments were conducted, in which two different pools of participants were each exposed to 80 IT trials. In each trial, participants had to indicate which of the legs (left or right) of the PI-shaped stimulus was longest, by pressing the left or right shift key. The independent variable was stimulus exposure time, which ranged from 14 to 153 ms. The results from Experiment 1 revealed that participant's blinking behavior was time-contingent, with participants blinking less when the stimulus was visible, as compared to before and after. Also, blinking during stimulus presentation correlated negatively with response accuracy. Furthermore, participants who experienced a brightness illusion had higher response accuracy as compared to others. Experiment 2 was a replication of Experiment 1 but featured enhanced task instructions and practice trials. Experiment 2 showed improved response accuracy, but no performance differences for the different illusions (or no illusion). In short, performance at the IT task is strongly affected by task familiarity and involves motor activity in the form of blinking.

Chapter 10: Discussion and conclusion

Chapter 10 discusses every important finding of the chapters, and places it in light of the questions that were asked in the Introduction. In short, the conclusions are as follows:

1. Wickens's SEEV model served as an excellent tool to structure and interpret the results of different chapters in the thesis. However, the use of perceptual heuristics by humans should be implemented in the model to create a more accurate representation of real-life gaze behavior.

2. Gaze behavior is indicative of task performance, especially for simpler tasks, like in the Senders replication experiment. Gaze behavior also revealed how people made use of visual feedback, which in itself improved task performance. However, measuring eye movements in itself does not necessarily reveal a connection between performance and the measurements.

3. Besides real-time performance prediction, this thesis also shows that eye-movements allow for a normative assessment for human-machine interface design.

VISUELE AANDACHT BIJ MENS–MACHINE INTERACTIE

Samenvatting

Mensen zijn niet in staat om overal tegelijkertijd aandacht aan te besteden. Het seriële karakter van gerichte aandacht beperkt de informatie-opnamecapaciteit van het perceptuele systeem.

Dit proefschrift gaat over het meten en modelleren van visuele aandachtsverdeling. Er wordt onderzocht of metingen van visuele aandacht voorspellend zijn voor de taakprestatie.

Hoofdstuk 1: Introductie

Het eerste hoofdstuk introduceert het hoofdonderwerp van dit proefschrift: de complexe aard van moderne technologische systemen, die veel informatiebronnen bevatten die gemonitord moeten worden.

In de literatuur over menselijke factoren zijn veel psychologische constructen voorgesteld die vermeende criteriumvaliditeit voor taakprestatie hebben. Hier wordt taakprestatie beschouwd als het vermogen van de mens om bijvoorbeeld de controle over een geautomatiseerd systeem over te nemen in mogelijk kritieke situaties. In tegenstelling tot de speculatieve aard van sommige van de Human-Factors constructen, tracht dit proefschrift prestaties vast te leggen in termen van objectieve metingen van visuele aandacht.

Hierna wordt het Salience, Effort, Expectancy, Value (SEEV) -model van Wickens (2008) geïntroduceerd en besproken. Dit model wordt gebruikt voor het interpreteren van de resultaten van eye-tracking. Ten slotte wordt een rationale gegeven voor de onderwerpen in het proefschrift. In hoofdstukken 2 tot en met 4 van dit proefschrift wordt het onderzoek van Senders (1983) in detail besproken door middel van replicatieonderzoek en een uitgebreide tutorial over zijn wiskundige modellen. Deze hoofdstukken bieden een empirische onderbouwing en conceptueel begrip van het concept van visuele aandacht. Hoofdstukken 5 tot en met 8 bespreken visuele aandacht in het licht van luchtverkeersleiding (ATC) en geautomatiseerd rijden, en worden beschouwd als geschikte casussen voor het meten van aandachtsverdeling en het voorspellen van taakprestaties. Hoofdstuk 9 onderzoekt taakprestatie en visuele aandacht binnen een psychometrische taak: Inspectie Tijd, een taak die een goede casus biedt voor het operationaliseren van het effect van aandacht op taakprestatie. Hoofdstuk 10 sluit af met een discussie over de onderwerpen in dit proefschrift.

Hoofdstukken 2-3: Visuele aandacht bij een dial-monitoring taak

Hoofdstukken 2 en 3 gaan in op het baanbrekende experimentele en theoretische werk van John Senders (1983) over de verdeling van visuele aandacht in een dial-monitoring

taak. In zijn onderzoek moesten deelnemers de overschrijdingen van drempelwaardes detecteren van klokjes die met verschillende snelheden (d.w.z. bandbreedtes) bewogen.

Hoofdstuk 2 richt zich op de experimentele replicatie van het werk van Senders, terwijl Hoofdstuk 3 de kwantitatieve wiskundige modellen van visual sampling, die Senders heeft voorgesteld, nader behandelt. In Hoofdstuk 2 werd aangetoond dat de oorspronkelijke resultaten van Senders uit 1964 met hoge nauwkeurigheid werden gerepliceerd. Bovendien werd in het replicatie-experiment aangetoond dat visual sampling niet alleen afhangt van de informatiebandbreedte van de stimulus (Verwachting), maar ook van de Opvallendheid van de stimulus en de Inspanning die nodig is om de volledige stimulus te overzien.

Hoofdstuk 3 interpreteert de resultaten van Hoofdstuk 2 op wiskundige wijze, en verklaart en verduidelijkt de originele wiskundige modellen van Senders. Bandbreedte-afhankelijke (dat wil zeggen, verwachtingsafhankelijke) sampling wordt beschreven door middel van het Periodic Sampling Model en het Random Constrained Sampling Model. Contextuele effecten, zoals Saliency en voorkennis van het signaal, worden meegenomen in het Conditional Sampling Model. Op basis van de Hoofdstukken 2 en 3 wordt geconcludeerd dat het werk van Senders een goede criteriumvaliditeit heeft voor het voorspellen van visual sampling in eenvoudige monitoringtaken.

Hoofdstuk 4: Naar het gebruik van visuele aandacht voor het meten van situatiebewustzijn

In lijn met het doel van dit proefschrift, namelijk het vinden van een metriek die taakprestatie voorspelt, biedt Hoofdstuk 4 een kritisch narratief over een baanbrekend construct in de Human Factors literatuur: Situation Awareness. Situation Awareness, een construct wat is geformaliseerd en geoperationaliseerd door Mica Endsley (1987, 1995), heeft criteriumvaliditeit voor taakprestatie in een breed spectrum van toepassingsgebieden (bijv. ATC, vliegen en autorijden). Hoofdstuk 4 breidt de discussie over het experiment van Hoofdstuk 2 uit, waarin we ook een veelgebruikte Situation Awareness meettechniek hebben toegepast: SAGAT. Geconcludeerd wordt dat SAGAT een bescheiden criteriumvaliditeit heeft voor taakprestatie. Oogbewegingen waren significant beter voorspellend voor de taakprestaties dan de SAGAT meting. Het hoofdstuk wordt afgesloten met een bespreking van de pragmatische toepassing van eye-tracking om Situation Awareness te meten.

Hoofdstukken 5-6: Visuele aandacht bij luchtverkeersleiding

Hoofdstukken 5 en 6 richten zich op taakprestatie en visuele aandachtsverdeling in de context van ATC.

Hoofdstuk 5 beschrijft een studie waarin deelnemers werden onderworpen aan een ATC-achtige conflictdetectietaak, waarbij ze continu moesten aangeven of twee

bewegende objecten zich op een ramkoers bevonden. De afhankelijke variabelen waren oogbewegingen en spatiebalk drukgedrag (d.w.z. conflictdetectie). Onafhankelijke variabelen waren de conflicthoek (30, 100, 150 graden), updatesnelheid (continu versus discreet) en de aanwezigheid van conflicten. De resultaten gaven aan dat hoeken van 30 graden de beste prestaties opleverden en 100 graden de slechtste. Bovendien leverden discrete stimuli een slechtere prestatie op dan stimuli met continue updatesnelheid. De betere prestaties op oppervlakkige conflicthoeken kunnen worden verklaard door perceptuele heuristieken, zoals de 'closer is first' strategie. Een oogbewegingsanalyse bevestigde deze op heuristiek gebaseerde hypothese voor ondiepe hoeken, aangezien deelnemers pursuit oogbewegingen gebruikten, terwijl deelnemers voor grotere conflicthoeken voornamelijk heen-en-weer-sampling gebruikten tussen vliegtuig en conflictpunt. Oogbewegingspatronen zijn dus voor een groot deel verklaarbaar in termen van de afstand tussen de punten, die groter is naarmate de conflicthoek groter is, een hypothese die in lijn is met het SEEV-model.

Hoofdstuk 6 onderzocht conflictdetectie prestaties, dit keer echter in een statische ATC-achtige taak. Daarin was het de bedoeling om het effect van augmented feedback uit te zoeken (een zogenaamd Solution Space Diagram; SSD) op de werklast, prestaties en visuele aandachtsverdeling van deelnemers. De resultaten gaven aan dat de augmented feedbackconditie in een lagere zelfgerapporteerde taakmoeilijkheid en een hoger conflictdetectiepercentage resulteerde. De percentages vals-positief waren ongeveer gelijk tussen de groepen. Bovendien besteedden de deelnemers aan de SSD-groep een groot deel van de tijd aan het kijken naar de SSD, ten koste van het kijken naar andere taakrelevante delen van de visuele scène.

Hoofdstukken 7-8: Visuele aandacht bij de perceptie van geautomatiseerde voertuigen

Hoofdstukken 7 en 8 hebben tot doel om taakprestaties en visuele aandachtsverdeling in de interactie tussen voetgangers en geautomatiseerde voertuigen (AV's) te operationaliseren. Beide hoofdstukken richten zich op het gebruik van externe mens-machine-interfaces (eHMI's) voor AV-voetganger communicatie.

Hoofdstuk 7 beschrijft een experiment waarin verschillende eHMI-plaatsingen werden geëvalueerd in termen van het spatiebalk drukgedrag van de deelnemers (als een index van wanneer deelnemers zich veilig voelden om over te steken) en oogbewegingen. De onafhankelijke variabele was de plaatsing van de eHMI (dak, voorruit, grill, boven de wielen of een projectie op de weg). De resultaten gaven aan dat wanneer de auto langzamer ging rijden, de eHMI's op het dak, de voorruit en de grill superieure prestaties leverden in vergelijking met de projectie en wielen eHMI's. Oogbewegingsanalyse lieten zien dat de projectie-eHMI meer verspreide oogbewegingen opleverde dan de andere eHMI's, wat aangeeft dat deelnemers meer heen en weer scanden tussen eHMI en

andere relevante elementen van de scène. Er werd geconcludeerd dat eHMI's aan verschillende kanten van het voertuig moeten worden gemonteerd voor optimale zichtbaarheid.

Hoofdstuk 8 gaat in op het zogenaamde message perspectief van de eHMI, met andere woorden, of een eHMI een instructieve boodschap moet weergeven (dwz 'WALK', of 'DON'T WALK', ook: egocentrische berichten) of een informatief bericht (dwz 'DRIVING' of 'BRAKING', ook wel: allocentrische berichten) om optimale communicatieve duidelijkheid te bereiken. Ook werd het effect van ambigue messages (d.w.z. 'STOP' en 'GO') onderzocht in termen van oogbewegingen en responsprestaties. Deelnemers werd gevraagd om met 'ja' of 'nee' (linker en rechter shift-toets) te reageren op de stelling 'ik kan oversteken' wanneer ze een foto te zien kregen van een auto met een eHMI waarop een van de bovengenoemde eHMI-messages werd weergegeven. Een geheugentaak werd geïncludeerd om het effect van real-life workload te simuleren. De resultaten van het experiment lieten zien dat egocentrische messages het meest overtuigend waren, wat blijkt uit meer consistente beslissingen om over te steken en snellere responstijden. Bovendien gaven de resultaten aan dat de ambigue messages werden geïnterpreteerd vanuit een egocentrisch perspectief, dat wil zeggen: 'GO' moedigde het oversteken aan en 'STOP' ontmoedigde het oversteken. Oogbewegingsanalyses lieten zien dat langere tekst een hoger aantal saccades veroorzaakten, maar de taakprestatie niet belemmerden. Geconcludeerd wordt dat eHMI's mogelijk egocentrische messages moeten weergeven.

Hoofdstuk 9: Visuele aandacht bij een psychometrische taak

Hoofdstuk 9 bestudeert visuele aandachtsverdeling in de context van een elementaire psychometrische taak: Inspectie Tijd (IT). Hier werd het effect van verschillende stimulusblootstellingstijden op taakprestaties onderzocht, gemeten aan de hand van responsnauwkeurigheid en responstijden. Verder werd het effect van hogere-orde strategieën en perceptie van visuele illusies geëvalueerd in de context van taakprestatie. Er werden twee experimenten met grote steekproeven uitgevoerd, waarbij twee verschillende groepen deelnemers elk werden blootgesteld aan 80 IT-stimuli. Bij elke stimulus moesten de deelnemers aangeven welke van de benen (links of rechts) van de PI-vormige stimulus het langst was, door op de linker of rechter shift-toets te drukken. De onafhankelijke variabele was de blootstellingstijd aan de stimulus, die varieerde van 14 tot 153 ms. De resultaten van Experiment 1 lieten zien dat het knippergedrag van de deelnemer tijdsafhankelijk was, waarbij deelnemers minder knipperen als de stimulus zichtbaar was, vergeleken met ervoor en erna. Knipperen tijdens stimuluspresentatie correleerde ook negatief met responsnauwkeurigheid. Bovendien hadden deelnemers die een helderheidsillusie ervoeren een hogere responsnauwkeurigheid in vergelijking met anderen. Experiment 2 was een replicatie van Experiment 1 maar bevatte verbeterde taakinstructies en oefenstimuli. Experiment 2 toonde een verbeterde

responsnauwkeurigheid, maar geen prestatieverschillen voor de verschillende illusies (of geen illusie). Kortom, prestaties bij de IT-taak worden sterk beïnvloed door taakbekendheid en bevat motorische activiteit in de vorm van knipperen.

Hoofdstuk 10: Discussie en conclusie

Hoofdstuk 10 bespreekt elke belangrijke bevinding van de hoofdstukken en plaatst deze in het licht van de vragen die in de inleiding werden gesteld. Samengevat zijn de conclusies als volgt:

1. Wickens' SEEV-model diende als een uitstekend hulpmiddel om de resultaten van verschillende hoofdstukken in het proefschrift te structureren en te interpreteren. Het gebruik van perceptuele heuristieken door mensen moet echter in het model worden geïmplementeerd om een nauwkeurigere weergave van het echte kijkgedrag te creëren.
2. Kijkgedrag is een indicatie van taakprestaties, vooral voor eenvoudigere taken, zoals in het replicatie-experiment van Senders. Kijkgedrag onthulde ook hoe mensen gebruik maakten van visuele feedback, wat op zichzelf de taakprestaties verbeterde. Het meten van oogbewegingen op zich hoeft echter niet per se een verband tussen prestatie en de metingen aan het licht te brengen.
3. Naast het real-time voorspelling van prestaties, laat dit proefschrift ook zien dat oogbewegingen een normatieve beoordeling mogelijk maken voor het ontwerp van mens-machine interfaces.

CHAPTER 1

Introduction

INTRODUCTION

Automation is found everywhere around us. From simple tasks, like washing the dishes to highly complex tasks, such as flying an airplane or driving a car, automation has entered every imaginable area of our lives. Generally, the implementation of automation has vast benefits over manual control. For example, technology in automated vehicles (AVs) has the potential to save many lives, as up to 95% of road accidents are caused by preventable human errors (e.g., Fagnant & Kockelman, 2015; NHTSA, 2016; ROSPA, 2017).

However, automation does not only bring moonshine and roses. A recent study by Meuller, Cicchino, and Zuby (2020) concluded that AVs may still make errors, even if they have perfect perception and show no incapacitation. They pointed out that high road traffic fatality rates could continue to persist due to, amongst others, AVs' errors in choosing evasive maneuvers, predicting the actions of other road users, and traveling at speeds unsuitable for the conditions (Meuller, Cicchino, & Zuby, 2020). The interaction between AVs and vulnerable road users (VRUs) appears to be an area with substantial implications for safety and traffic efficiency (Millard-Ball, 2018).

Contemporary on-road AVs only feature SAE level 2 automation (e.g., Tesla's Autopilot), or in specific cases, an extended version of level 3 automation, dubbed level 2+ automation (Nvidia, 2020[1]) that adds, amongst other things, basic driver monitoring. Level 2 automated driving, or partial automation[2], means that the vehicle is capable of automatic acceleration, deceleration, and can perform certain steering maneuvers. However, human drivers still need to be able to take over control in case the automation does not perform safely, for example when a pedestrian steps onto the road (e.g., Gold et al., 2016; Petermeijer et al., 2017). This readiness to take over control requires that the driver has sufficient awareness (De Winter et al., 2018; Endsley, 1995) of the environment and sufficient knowledge of the automation systems. Recent accidents with level 2 AVs demonstrate that the driver is not always capable of taking over control (Wikipedia, 2020[3]). A recent news article reported an extreme case in which the 'driver' of a Tesla apparently fell asleep[4] while being in autopilot mode. This case, and many other cases in which drivers failed to monitor the automation, are examples of the adverse effects of automation on human operators.

1 https://blogs.nvidia.com/blog/2019/02/06/what-is-level-2-automated-driving/
2 https://www.synopsys.com/automotive/autonomous-driving-levels.html#:~:text=Level%202%20(Partial%20Driving%20Automation,the%20car%20at%20any%20time.
3 https://en.wikipedia.org/wiki/List_of_self-driving_car_fatalities
4 https://arstechnica.com/tech-policy/2019/09/how-tesla-could-fix-its-sleeping-driver-problem/

Automation monitoring

Long before automated cars existed, Mackworth (1948) demonstrated with his famous Mackworth's clock experiment that sustained attention on a simple monitoring task resulted in a significant performance decrement after only 10 minutes, a result that has been replicated many times (e.g., Lichstein, Riedel, & Richman, 2000). In the context of automation monitoring, Bainbridge (1983) stated that "it is humanly impossible to carry out the basic function of monitoring for unlikely abnormalities […]" (p. 776) for extended periods, "which therefore has to be done by an automatic alarm system connected to sound signals" (ibid. p.776). Parasuraman and Riley (1997) added to this narrative by pointing out the risks of misuse of automation. Here, the operator assumes that the automation is more capable than it actually is and uses it in a way the automation was never designed for, for example, by using his or her phone too much. These observations do not only apply to semi-autonomous driving but also to a plethora of other (partially) automated tasks. The aforementioned issues with the monitoring of automation are relatively new because automated systems (such as AVs) have become available to the general public recently.

In order to create safe interactions with automated systems, the automation must have an idea about whether the human is able to take over control, or whether the human needs support and assistance. To counteract the abovementioned "ironies of automation" (Bainbridge, 1986), there is a need for an empirical measure that has criterion validity with regard to human ability and performance in taking over control.

Contrary to automated systems, manual controlled systems allow for a relatively straightforward measurement of driver engagement and task performance. For example, in driving, a simple measure like the standard deviation of lateral position could be used as proxy and predictor for driver engagement. For automated systems, where the operator is not physically engaging with the system anymore, online measurement and prediction of task performance is not so easy. This thesis sets out to find measures and models of task performance that allow for online and continuous measurement in automated systems.

Within the domain of Human Factors and Ergonomics, numerous psychological constructs have been proposed that have alleged predictive validity with regards to task performance: Situation Awareness (SA; Endsley, 1988, 1995; Smith & Hancock, 1995), workload (Hart & Staveland, 1988; Wickens, 2008 and De Waard, 1996), mode awareness (e.g., Sarter & Woods, 1995; 1995; Kurpiers et al. 2020), and many more (see also Heikoop et al., 2015). There has been a vigorous debate in the literature as to whether these constructs are scientifically credible (Dekker and Woods, 2002; Dekker and Hollnagel, 2004; Dekker et al. 2010, Parasuraman, Sheridan & Wickens, 2008) and operationalizable (De Winter, 2014; Sarter & Woods, 1991; Salmon et al.,

2009). Moreover, even though certain Human Factors constructs correlate with task performance, they often are measured offline and discretely in time (e.g., query items in case of the widely used SAGAT method for measuring SA, Endsley; 1995) and are not suitable for real-time, online measurements.

Overt visual Attention

As an alternative to the aforementioned constructs, overt visual attention has been proposed as an online and continuous measure that reflects the operator's awareness of the situation and the monitored system, and accordingly may be predictive of task performance (Senders et al., 1967; Moore & Gugerty, 2010; Van de Merwe, Van Dijk, & Zon, 2012). The use of visual attention measures is grounded on the assumption that the information at which one fixates is likely to be the subject of concurrent cognitive processes. This hypothesis has been formalized by Just and Carpenter (1976, 1980) in the so-called 'eye-mind assumption' (or eye-mind hypothesis), stating that "there is no appreciable lag between what is being fixated and what is being processed" (Just & Carpenter, 1980; p. 331), as well as the 'immediacy assumption', which indicates that "the interpretations at all levels of processing are not deferred; they occur as soon as possible" (Just & Carpenter, 1980; p. 330).

In line with the eye-mind assumption, this thesis only considers the foveal aspect of visual attention. In general, the visual field consist of two major components: the sharp foveal part and the blurred peripheral part. During a fixation, the foveal part of the retina is directed towards the point of fixation. In line with the eye-mind assumption, I assume that the majority of the information transfer occurs at the point of fixation, at the location where one fixates at. Of course, the effect of peripheral vision on information transfer is not to be neglected. Salient events in the visual periphery may be the cause of eye-movements towards a certain point of interest in the visual scene, thereby changing the point of fixation. However, the focus on events in the periphery is essentially implicit (covert); that is: no overt eye-movements are needed to shift attention towards other parts of the peripheral field. To measure the effect of covert visual attention, or in other words: to operationalize the effect of peripheral vision, one could make use of so-called gaze-contingent windows. In this thesis, I am first and foremost interested in the relation between overt visual attention and task performance, for two reasons: (1) because of the theoretical assumption that is made in the eye-mind hypothesis, namely that the information content on the location of fixation is likely to be the prime constituent of immediate cognitive processes (also in line with the immediacy assumption), (2) because overt visual attention can be easily measured by means of eye-tracking and allows for experimental control, which stands in contrast with the delicate experimental control one must have to operationalize the effect of covert visual attention.

Overt visual attention distribution is a purported predictor of task performance, correlating with performance measures, such as reaction times and task accuracy (Gegenfurtner, Lehtinen & Saljo, 2011; Reingold & Sheridan, 2011). Pioneering research regarding the effect of visual attention distribution on task performance has been done in the area of chess (De Groot, 1946; Chase & Simon, 1973,b). In Chase and Simon (1973), chess grandmasters were able to accurately (93% correct positions for 25 pieces) remember complex board configurations after only short exposure times (2–15 seconds), outperforming less skilled players to a great extent. These findings suggest that grandmasters do not rely on (slow) serial visual scanning of the chessboard (which novices do) but rather use a different process that Chase and Simon termed 'chunking'; fast parallel encoding of meaningful structures on the chessboard. Reingold et al. (2001) verified these results and showed that experts exhibited fewer saccades and fixations while maintaining superior performance (see also Charness et al., 2001). The authors concluded: "The present study illustrates that eye movement paradigms may prove invaluable in supplementing traditional measures of performance such as RT, accuracy, and verbal reports as a means for understanding human expertise in general and chess skill in particular." (Reingold et al. 2001; p. 55). Thus, by understanding how visual attention is distributed in the context of a task, one could gain a quantitative understanding of how humans will perform in that task.

In summary, a measure of overt visual attention is likely to be a good candidate for predicting task performance. In literature, overt visual attention is often equated with the term 'eye movements' and is traditionally measured using eye trackers (i.e., Yarbus, 1967; Senders, 1964). However, when endeavoring to learn what has been written on visual attention and how it could be modelled and measured, one is inclined to lose hope quickly. "Everyone knows what attention is […]" (p. 381) said William James in his Principles of Psychology, namely: "[…] it implies withdrawal from some things in order to deal effectively with others" (p. 382). However, a Google Scholar search[5] on the term "Visual Attention" reveals otherwise. No less than 4.06 million search results demonstrate the plethora of views and perspectives on the concept of visual attention. The indexed articles cover topics from the underlying neural mechanisms of visual attention (e.g., Koch & Ullman, 1987; Desimone & Duncan, 1995), shared visual attention in infants (e.g., Scaife & Bruner, 1975), saliency based modelling of visual attention (e.g., Itti, Koch, & Niebur, 1998; Itti & Koch, 2000), distraction of visual attention due to, for example, cellphone use in car driving (Strayer, Drews, & Johnston, 2003, Strayer & Drews, 2007) to visual search models (Wolfe, 1994, 2010).

5 https://scholar.google.nl/scholar?hl=nl&as_sdt=0%2C5&q=Visual+Attention&btnG=

From the spectrum of views on visual attention, I have chosen to use the Salience-Effort-Expectancy-Value (SEEV) model of Wickens and colleagues (Wickens, 2008) as a structural framework. I use this framework to understand what factors influence the distribution of visual attention and to shape the results of this thesis as well as the corresponding discussions and conclusion. As explained below, the SEEV model encompasses the factors that, based on a wealth of literature, are known to be predictive of the probability that a human looks at a particular element of the visual field.

Wickens's (2008) SEEV model

SEEV is a qualitative model (and to some extent transformable in a quantitative model, see e.g., Steelman, McCarley, & Wickens, 2011; Bundesen, 1990) that aims to describe the distribution of overt visual attention on the basis of two bottom-up and two top-down factors. In Wickens's (2008b) words, SEEV is a model that identifies the parameters that "drive the eyeball (visual attention) around the environment" (p. 5). Saliency and Effort are the two bottom-up factors that direct and limit the movement of the attentional spotlight across the visual scene. Saliency, defined by Wickens (2008b[6]) as "the bottom-up attention capturing properties of events, bright flashes, sounds, etc." (p. 5) represents the bottom-up mechanism of *attracting* attention, and Effort denotes a bottom-up *limitation* on the distribution of attention. In other words, Effort "inhibits the movement of attention across longer distances: bigger scans, head movements." (p.5).

Expectancy and Value are the so-called top-down factors in the model. Expectancy-driven distribution of attention is facilitated by the mental model an individual has of a certain situation. Expectancy is based on the "likelihood of seeing an event at a particular location […]" (p.5). Value is the second top-down factor in the SEEV model, and it represents the importance of (not) attending to a certain event in the visual scene, as well as the "relevance of the event to a valued task" (p.5). In Figure 1, a schematic representation of the SEEV model is provided. Here, the input of the models consists of four factors, and the output of the model is represented by the probability that someone will attend to a certain area of the visual scene. This can be expressed as follows:

$$P(A_i) = \boldsymbol{S}_i - \boldsymbol{E}f_i + \boldsymbol{E}x_i + \boldsymbol{V}_i$$

Here, $P(A_i)$ stands for the probability that an observer will sample a specific area of interest in the visual scene, and the subscript *i* denotes the number of the area. The term "areas of interest" refers to areas of the visual scene that are of potential relevance to the task at hand.

6 Presentation Wickens (2008b)

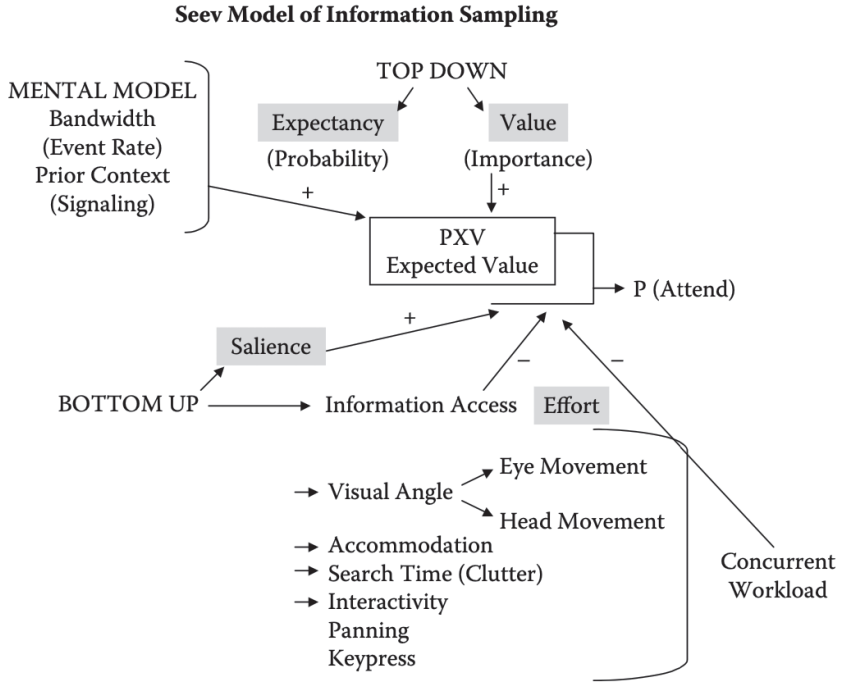


Figure 1. SEEV model per Wickens (2008b). This schematic representation of the SEEV relates all the factors of the SEEV-model in a visual way.

So far, it seems that the SEEV model is solely descriptive in nature; however, it may also be used in a more normative manner. For example, in case of designing a human-machine interface (e.g., a car dashboard or aircraft cockpit), the designer could utilize the SEEV model to optimize the probability that the operator will attend to a certain part of the interface at some point in time, or, in Wickens' (2008b) words, to "make valuable information Salient" and "reduce the Effort of transitioning between sources with high Expectancy or bandwidth" (p. 6). In terms of the mathematical operationalization of the model, this design strategy could be expressed as follows: to optimize $P(A_i)$ for some interface design, one should minimize the inherent distracting factors (which are bottom-up) denoted by the term $\{\boldsymbol{S}_i - \boldsymbol{Ef}_i\}$, and maximize the probability of top-down-based sampling, as denoted by the term $\{\boldsymbol{Ex}_i + \boldsymbol{V}_i\}$**.** The SEEV model has been successfully used in a study by Steelman, McCarley, and Wickens (2011) to predict the distribution of visual attention of pilots in different phases of the flight. More specifically, they predicted the observed percentage dwell times (PDT) with high accuracy (R^2 = 0.9).

Scopes and structure of the thesis

As pointed out above, there is a need for real-time measurements that have criterion validity with regard to the operator's task performance. In this thesis, I propose that eye movements may contribute to such real-time assessment. Accordingly, this thesis presents a number of articles in which the relationship between eye movements and task performance is studied in a theoretical and empirical manner. The main question that this thesis examines is: "can we use visual attention as a proxy to predict and explain task performance?" In other words, this thesis examines (1) whether visual attention can predict task performance, and explores (2) which task-related factors drive "the eyeball", using the SEEV model as an interpretative framework.

This thesis is focused on three main application areas: (1) Air Traffic Control (ATC), a safety-critical domain in which operators, amongst other tasks, need to visually identify conflicts between aircraft, (2) automated driving, also a safety-critical domain, in which AVs need to visually communicate with vulnerable road users, and (3) psychometrics, in which we take a fundamental approach towards evaluating task performance in the (alleged) simplest psychological task that exists: Inspection Time (IT). The work that is presented in this thesis is the result of a collaboration between the faculties of Mechanical Engineering (3mE) and Aerospace Engineering (AE), which is reflected in the choice of topics. I conducted the research on ATC at the faculty of AE, whereas the driving-related and fundamental psychometrics research were conducted at 3mE. Both ATC and automated driving are topics that are suitable use cases for attention distribution modelling and measurement. Research on basic psychometric tasks provides a good testbed for operationalizing the effect of attention on task performance.

From a historical and contemporary perspective, Air Traffic Control (e.g., Fitts, 1951) and (automated) driving are researched extensively in Human Factors. For both of these areas, various levels and stages of automation (e.g., from advisory Human Machine Interfaces to automated control) have been introduced to enhance safety and the operators' (or drivers') task performance (see Fitts, 1951). For the case of automated driving, it is often argued that automated vehicles behave differently as compared to manual driven vehicles, and that traditional ways of communication between the AVs 'driver' and other road users (e.g., eye-contact and gesturing) are disappearing. Displays on the outside of the car (also called: External Human Machine Interfaces, or eHMIs) have been introduced to reinstate these disappearing modes of communication. However, there exists no consensus in the literature as to how (augmented) feedback should be designed, both in case of ATC and eHMIs, and what the effect of this feedback is on task performance. We have performed several eye-tracking experiments in the two application areas to quantify the (limiting) effect of visual attention on participant's

task performance and to identify the potential up – and downsides of automation and feedback in both areas.

This thesis comprises four main parts, followed by a separate chapter with discussion and conclusion.

1. **Introduction and theoretical background on Attention.** In this part of the thesis, a theoretical introduction of visual attention is provided, specifically in the context of Situation Awareness and Senders's (1983) quantitative models of visual attention. In Chapter 2, Senders's (1983) six dial experiment is replicated with the aim of identifying gaze-directing factors (in line with SEEV). In Chapter 3, Senders's (1983) quantitative models of visual sampling are explained and discussed further. Chapter 4 discusses the inherent problems with Endsley's (1995) conceptualization of SA, and a new eye-movement-based SA construct is proposed.
2. **Application area 1: Visual attention distribution in Air Traffic Control.** In Chapter 5, we focus on evaluating the effect of attention distribution in a dynamic allocentric (ATC) conflict detection task. In Chapter 6, we aim to quantify the mediating effect of (selective) visual attention on performance in an ATC-like static conflict-detection task. The effect of a novel ATC feedback tool (called the SSD) is evaluated in terms of eye-movements and task performance.
3. **Application area 2: Visual attention distribution in AV-pedestrian interaction.** Chapters 7 and 8 aim at quantifying visual attention distribution in vehicle-pedestrian interactions, with a special focus on evaluating the potential benefits of eHMIs. The main focus of Chapter 7 is the experimental evaluation of eHMI position on pedestrian crossing behavior, whereas Chapter 8 focusses on researching the effect of eHMI message perspective on pedestrian crossing decisions.
4. **Application area 3: Attention distribution in psychometrics.** Chapter 9 studies the effect of sustained attention on Inspection Time (IT) performance. IT performance has been extensively used as correlational proxy for intelligence (g), and it is the purpose of this last chapter to research whether attention is a mediating variable for task performance on the IT task.
5. **Conclusion and discussion.** Chapter 10 gives a summary of the results of the thesis and answers the questions that are posed in the introduction.

In light of the aim of this thesis, which is to develop a real-time metric of visual attention for predicting task performance, each and every one of the forthcoming experiments feature real-time and high frequent measurement of eye-movements. The diverse nature of the experiments that are described here, allow for creating a multi-facetted perspective on real-time operator assessment, on the one hand from an abstract

perspective (e.g. as Chapter 3 and 9) and on the other hand from a more ecological viewpoint (e.g. Chapter 7). As mentioned earlier, the discussion is structured along the lines of the SEEV model, and is intended to identify the constituent factors that drive the eyeball over the visual scene, as a function of the operator's cognitive processes (top-down) as well as factors from the environment (bottom-up). The identification and quantification of these factors (see for example the experiment carried out in Chapter 2) consequently attribute to a better understanding of how to assess the operator in the context of the task that is being carried out. The long-term outlook ultimately comprises utilizing the developed understanding of visual attention, and attention in general, for real-time operator assessment.

REFERENCES

Bainbridge, L. (1983). Ironies of automation. *Automatica, 19*(6), 775-779. doi:10.1016/0005-1098(83)90046-8

Bundesen, C. (1990). A theory of visual attention. *Psychological Review*, *97*(4), 523–547. https://doi.org/10.1037/0033-295x.97.4.523

Charness, N., Reingold, E. M., Pomplun, M., & Stampe, D. M. (2001). The perceptual aspect of skilled performance in chess: Evidence from eye movements. *Memory & Cognition, 29*(8), 1146-1152. doi:10.3758/bf03206384

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4*(1), 55-81. doi:10.1016/0010-0285(73)90004-2

Dekker, S. W., & Woods, D. D. (2002). MABA-MABA or Abracadabra? Progress on human-automation coordination. *Cognition, Technology & Work, 4*(4), 240-244. doi:10.1007/s101110200022

Dekker, S. W., Nyce, J. M., Winsen, R. V., & Henriqson, E. (2010). Epistemological self-confidence in human factors research. *Journal of Cognitive Engineering and Decision Making, 4*(1), 27-38. doi:10.1518/155534310x495573

Dekker, S., & Hollnagel, E. (2004). Human factors and folk models. *Cognition, Technology & Work, 6*(2), 79-86. doi:10.1007/s10111-003-0136-9

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18*(1), 193-222. doi:10.1146/annurev.ne.18.030195.001205

Endsley, M.R. (1988). Situation awareness global assessment technique (SAGAT). *Proceedings of the IEEE 1988 National Aerospace and Electronics Conference*. doi:10.1109/naecon.1988.195097

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 37*(1), 32-64. doi:10.1518/001872095779049543

Fagnant, D. J., & Kockelman, K. (2015). Preparing a nation for autonomous vehicles: Opportunities, barriers and policy recommendations. *Transportation Research Part A: Policy and Practice, 77*, 167-181. doi:10.1016/j.tra.2015.04.003

Fitts, P. M. (Ed.). (1951). *Human engineering for an effective air-navigation and traffic-control system.* National Research Council, Div. of.

Gegenfurtner, A., Lehtinen, E., & Säljö, R. (2011). Expertise differences in the comprehension of visualizations: a meta-analysis of eye-tracking research in professional domains. *Educational Psychology Review, 23*(4), 523-552. doi:10.1007/s10648-011-9174-7

Gold, C., Körber, M., Lechner, D., & Bengler, K. (2016). Taking over control from highly automated vehicles in complex traffic situations. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 58*(4), 642-652. doi:10.1177/0018720816634226

Groot, A. D. (1946). *Het denken van den schaker, een experimenteel-psychologische studie*. Noord-Hollandsche Uitgevers Maatschappij, Amsterdam

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. *Advances in Psychology Human Mental Workload,* 139-183. doi:10.1016/s0166-4115(08)62386-9

Heikoop, D. D., Winter, J. C., Arem, B. V., & Stanton, N. A. (2015). Psychological constructs in driving automation: A consensus model and critical comment on construct proliferation. *Theoretical Issues in Ergonomics Science, 17*(3), 284-303. doi:10.1080/1463922x.2015.1101507

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40*(10-12), 1489-1506. doi:10.1016/s0042-6989(99)00163-7

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(11), 1254-1259. doi:10.1109/34.730558

James, W. (1890). *The principles of psychology in two volumes*. New York: Holt.

Just, M. A., & Carpenter, P. A. (1976). The role of eye-fixation research in cognitive psychology. *Behavior Research Methods & Instrumentation, 8*(2), 139-143. doi:10.3758/bf03201761

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: from eye fixations to comprehension. *Psychological Review, 87*(4), 329-354. doi:10.1037/0033-295x.87.4.329

Koch, C., & Ullman, S. (1987). Shifts in selective visual attention: towards the underlying neural circuitry. *Matters of Intelligence,* 115-141. doi:10.1007/978-94-009-3833-5_5

Kurpiers, C., Biebl, B., Hernandez, J. M., & Raisch, F. (2020). Mode awareness and automated driving—what is it and how can it be measured? *Information, 11*(5), 277. doi:10.3390/info11050277

Lichstein, K. L., Riedel, B. W., & Richman, S. L. (2000). The Mackworth clock test: a computerized version. *The Journal of Psychology, 134*(2), 153-161. doi:10.1080/00223980009600858

Mackworth, N. H. (1948). The breakdown of vigilance during prolonged visual search. *Quarterly Journal of Experimental Psychology, 1*(1), 6-21. doi:10.1080/17470214808416738

Merwe, K. V., Dijk, H. V., & Zon, R. (2012). Eye movements as an indicator of situation awareness in a flight simulator experiment. *The International Journal of Aviation Psychology, 22*(1), 78-95. doi:10.1080/10508414.2012.635129

Moore, K., & Gugerty, L. (2010). Development of a novel measure of situation awareness: the case for eye movement analysis. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 54*(19), 1650-1654. doi:10.1177/154193121005401961

Mueller A.S., Cicchino J.B., Zuby D.S., (2020). What humanlike errors do autonomous vehicles need to avoid to maximize safety? Arlington (VA): Insurance Institute for Highway Safety. https://iihs.org.

Parasuraman, R., & Riley, V. (1997). Humans and automation: use, misuse, disuse, abuse. *Human Factors, 39*, 230-253. doi:https://doi.org/10.1518/001872097778543886

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2008). Situation awareness, mental workload, and trust in automation: viable, empirically supported cognitive engineering constructs. *Journal of Cognitive Engineering and Decision Making, 2*(2), 140-160. doi:10.1518/155534308x284417

Petermeijer, S., Bazilinskyy, P., Bengler, K., & Winter, J. D. (2017). Take-over again: Investigating multimodal and directional TORs to get the driver back into the loop. *Applied Ergonomics, 62*, 204-215. doi:10.1016/j.apergo.2017.02.023

Reingold, E. M., & Sheridan, H. (2011). Eye movements and visual expertise in chess and medicine. *Oxford Handbooks Online*. doi:10.1093/oxfordhb/9780199539789.013.0029

Reingold, E. M., Charness, N., Pomplun, M., & Stampe, D. M. (2001). Visual span in expert chess players: evidence from eye movements. *Psychological Science, 12*(1), 48-55. doi:10.1111/1467-9280.00309

Salmon, P. M., Stanton, N. A., Walker, G. H., Jenkins, D., Ladva, D., Rafferty, L., & Young, M. (2009). Measuring situation awareness in complex systems: comparison of measures study. *International Journal of Industrial Ergonomics, 39*(3), 490-500. doi:10.1016/j.ergon.2008.10.010

Sarter, N. B., & Woods, D. D. (1991). Situation awareness: a critical but ill-defined phenomenon. *The International Journal of Aviation Psychology, 1*(1), 45-57. doi:10.1207/s15327108ijap0101_4

Sarter, N. B., & Woods, D. D. (1995). Autonomy, authority, and observability: properties of advanced automation and their impact on human-machine coordination. *IFAC Proceedings Volumes, 28*(15), 149-152. doi:10.1016/s1474-6670(17)45224-4

Sarter, N. B., & Woods, D. D. (1995). How in the world did we ever get into that mode? Mode Error and Awareness in Supervisory Control. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 37*(1), 5-19. doi:10.1518/001872095779049516

Scaife, M., & Bruner, J. S. (1975). The capacity for joint visual attention in the infant. *Nature, 253*(5489), 265-266. doi:10.1038/253265a0

Senders, J. (1964). The human operator as a monitor and controller of multidegree of freedom systems. *IEEE Transactions on Human Factors in Electronics, HFE-5*(1), 2-5. doi:10.1109/thfe.1964.231647

Senders, J. W. (1983). *Visual sampling processes* (Unpublished master's thesis). Tilburg.

Senders, J., Kristofferson, A., Levison, W., Dietrich, C., & Ward, J. (1967). The attentional demand of automobile driving. *Highway Research Record, 195*, 15-33.

Smith, K., & Hancock, P. A. (1995). Situation awareness is adaptive, externally directed consciousness. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 37*(1), 137-148. doi:10.1518/001872095779049444

Steelman, K. S., Mccarley, J. S., & Wickens, C. D. (2011). Modeling the control of attention in visual workspaces. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 53*(2), 142-153. doi:10.1177/0018720811404026

Strayer, D. L., & Drews, F. A. (2007). Cell phone–induced driver distraction. *Current Directions in Psychological Science, 16*(3), 128-131. doi:10.1111/j.1467-8721.2007.00489.x

Strayer, D. L., Drews, F. A., & Johnston, W. A. (2003). Cell phone-induced failures of visual attention during simulated driving. *Journal of Experimental Psychology: Applied, 9*(1), 23-32. doi:10.1037/1076-898x.9.1.23

Waard, D. D. (1996). *The measurement of drivers' mental workload.* Groningen: Traffic Research Centre, Univ. of Groningen.

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 50*(3), 449-455. doi:10.1518/001872008x288394

Wickens, C., & McCarley, J. (2008). *Applied Attention Theory.* Boco-Ratan, FL: CRC Press, Taylor & Francis.

Winter, J. C. (2014). Controversy in human factors constructs and the explosive use of the NASA-TLX: A measurement perspective. *Cognition, Technology & Work, 16*(3), 289-297. doi:10.1007/s10111-014-0275-1

Winter, J. C., Eisma, Y. B., Cabrall, C. D., Hancock, P. A., & Stanton, N. A. (2018). Situation awareness based on eye movements in relation to the task environment. *Cognition, Technology & Work, 21*(1), 99-111. doi:10.1007/s10111-018-0527-6

Wolfe, J. M. (1994). Guided Search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review, 1*(2), 202-238. doi:10.3758/bf03200774

Wolfe, J. M. (2010). Guided Search 4.0: A guided search model that does not require memory for rejected distractors. *Journal of Vision, 1*(3), 349-349. doi:10.1167/1.3.349

Yarbus, A. L. (1967). Eye movements during perception of complex objects. *Eye Movements and Vision,* 171-211. doi:10.1007/978-1-4899-5379-7_8

CHAPTER 2

Visual Sampling Processes Revisited: Replicating and Extending Senders (1983) Using Modern Eye-Tracking Equipment

Eisma, Y. B., Cabrall, C. D. D., & De Winter, J. C. F. (2018). Visual sampling processes revisited: Replicating and extending Senders (1983) using modern eye-tracking equipment. *IEEE Transactions on Human Machine Systems, 48, 526–540*.
Joint first authors

ABSTRACT

In pioneering work, Senders (1983) tasked five participants to watch a bank of six dials, and found that glance rates and times glanced at dials increase linearly as a function of the frequency bandwidth of the dial's pointer. Senders did not record the angle of the pointers synchronously with eye movements, and so could not assess participants' visual sampling behavior in regard to the pointer state. Because the study of Senders has been influential but never repeated, we replicated and extended it by assessing the relationship between visual sampling and pointer state, using modern eye-tracking equipment. Eye tracking was performed with 86 participants who watched seven 90-second videos, each video showing six dials with moving pointers. Participants had to press the spacebar when any of the six pointers crossed a threshold. Our results showed a close resemblance to Senders' original results. Additionally, we found that participants did not behave in accordance with a periodic sampling model, but rather were conditional samplers, in that the probability of looking at a dial was contingent on pointer angle and velocity. Finally, we found that participants sampled more in agreement with Nyquist sampling when the high bandwidth dials were placed in the middle of the bank rather than at its outer edges. We observed results consistent with the saliency, effort, expectancy, and value model and conclude that human sampling of multidegree of freedom systems should not only be modeled in terms of bandwidth but also in terms of saliency and effort.

1. INTRODUCTION

Technological systems are automated to ever greater extents [1]. In many automated systems, the role of the human is to monitor the instruments in order to assess whether the automation performs satisfactorily [2]. Present-day automated systems, such as aircraft cockpit, produce much more information than a human can process at once [3]. Consequently, the human needs to distribute attention across multiple sources of information in order to maintain accurate awareness of the automation state.

How humans sample dynamic instruments is a question that has been of broad interest in human factors and ergonomics (e.g., [4], [5]). Especially in the aviation domain, several studies have been performed that investigated how pilots distribute their visual attention across the different instruments in the cockpit [6]–[9]. In a seminal study, Fitts et al. [10] examined how 40 pilots distributed visual attention across cockpit instruments during aircraft landings. Based on their findings, Fitts et al. [10] argued that the number of eye fixations per second on an instrument is a measure of the importance of that instrument for carrying out the flight task. Additionally, the fixation duration on the instrument was regarded as an index of the difficulty in reading and interpreting the particular instrument. As pointed out by Landry [11] and Seeberger and Wierwille [12], the results of Fitts et al. [10] have been used to redesign the default lay-out of the cockpit instrument panel in that the instruments most frequently looked at are placed in the middle of the instrument cluster.

Further pioneering work on human sampling behavior of instruments was carried out by Senders [13]. He used the Nyquist-Shannon sampling theorem [14] to predict how frequently a human needs to sample an instrument in order to keep track of its state. This theorem can be intuitively understood when trying to reconstruct a sine wave from a number of periodically sampled data points of this sine wave. If not sampling with at least twice the frequency of the sine wave, then the sine wave cannot be reconstructed from those data points. Accordingly, Senders [15] postulated that if an instrument provides information with a frequency bandwidth W , the human as a Nyquist sampler (ideal observer) should observe that signal with a frequency equal or greater than 2W.

To test his theory, Senders [13] conducted an experiment in which five undergraduate students monitored a bank of four circular dials (microammeters), with randomly moving pointers that differed in bandwidth (0.08, 0.16, 0.32, and 0.64 Hz). The participants were instructed to press a response key (see Fig. 1) each time one of the four pointers crossed a threshold value from either side. They performed this monitoring task for one hour per day for 30 days. A 3-minute data sample of camera recordings pointed at the eyes of the subjects was collected and analyzed per hour of monitoring.

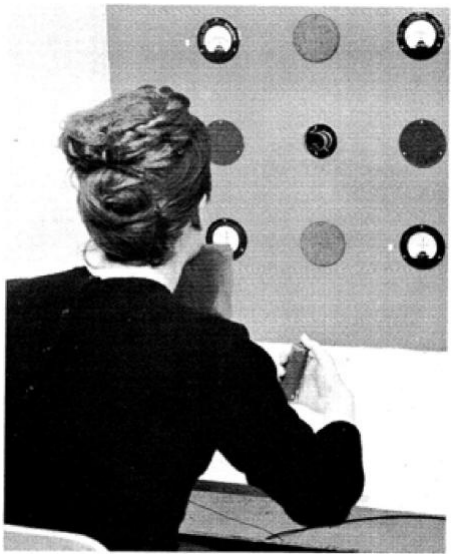


Figure 1. Illustration of one of the participants in a four-dial sampling task (photo from [17]). A motion picture camera is located in the middle of the four dials. The participant holds a switch in her right hand.

The results revealed a strong linear relationship between the signal bandwidth (W) and the average observed glance rate (GR) per dial (GR = 0.05 + 2.44 W, r = 0.98), offering clear support for Senders' theory. Moray [16] suggested that, because eye movements are so strongly predicted by signal bandwidth (r = 0.98), Fitts et al. [10] may have been mistaken in that not the importance (e.g., value, cost of missing) of an instrument, but rather its experienced bandwidth (i.e., expectancy) is the prime determinant of how frequently the human looks at an instrument. Put simply, it is possible that pilots in Fitts et al. looked at particular instruments more often than at other instruments not necessarily because these instruments were important for the flight task, but rather because these instruments had fast-moving pointers. However, this hypothesis could not be tested because the actual values of the instrument pointers were not recorded by Fitts and co-workers.

In his Ph.D. thesis published almost 20 years later, Senders [15] presented the results of four additional experiments also carried out in the 1960s [17]. These additional experiments were performed using five high school students who viewed six dials of different bandwidths (0.03, 0.05, 0.12, 0.20, 0.32, and 0.48 Hz). These four experiments were similar to each other, but differed somewhat in composition (i.e., a baseline experiment was performed, in a second experiment the random signals were generated in a slightly different manner, in a third experiment a binary signal was used for the 0.12 Hz dial, and in a fourth experiment the bandwidths were slightly varied). Participants received extensive training of at least 10 h. The results of the four

aggregated experiments again yielded a nearly perfect linear relationship between bandwidth and glance rate (r = 0.99), but with a slope that was considerably shallower (GR = 0.18 + 0.61W) than predicted by the Nyquist-Shannon theorem and Senders' 1964 experiment (2.44W). Relative to the model predictions, the shallower slope indicates that participants oversampled the low bandwidth dials while undersampling the high bandwidth dials. One explanation for the undersampling could be that participants tended to forget the state of the low-frequency signals [15], [16]. Furthermore, according to Senders, the introduction of the two very low bandwidth signals (0.03 and 0.05 Hz) may have increased the demands on participants to memorize the state of these dials, in turn causing them to pay less attention to the high bandwidth dials.

Another explanation for a slope shallower than 2W is the notion that participants may have been able to read the angular velocity of the pointers in addition to the pointers' current angle. This may have reduced the required sampling frequency from 2W to W [15], [18]. This extension of the sampling theorem can again be intuitively understood by trying to reconstruct a sine wave. If periodically sampling data points of the sine wave, plus the slope of said data points, then the original sine wave can be reconstructed when sampling only once per ordinary frequency of the sine wave. However, the extended sampling theorem cannot explain the different slopes found between Senders' four dial and six dial experiments.

Senders [13], [15] noted that although humans sample in accordance with a periodic sampling model (for the four dial configuration), it is unlikely that humans are actually periodic samplers who deterministically reconstruct a signal according to the sampling theorem, and who do not adjust their sampling behavior based on the momentary state of the pointers. In his thesis, Senders [15] proposed a number of "conditional sampling" models that predict the probability of sampling a particular dial as a function of the current state of the dial relative to the threshold, rather than its overall stochastic property (i.e., bandwidth). Moray [16] eloquently explained why conditional sampling models are viable: "suppose that an observation shows that the function is very close to the permissible limit. It seems likely that another fixation on that source would be made sooner than if it had been observed at, say, its mean" (pp. 40:11). However, because the technology of the 1960s did not allow for a synchronized recording of eye movements and the state of the dials, it still remained to be tested whether conditional models are more valid than a periodic sampling model that uses bandwidth as input. As noted by Senders [15]: "It is necessary to record not only the positions of the eyes but also the value of the signals which are observed. It is only the relationship of these two sets of data that will tell us whether there is anything at all in the idea that observers make use of the information that they see in deciding when to look again." (p. 98).

Various other researchers have proposed conditional models of visual sampling. For example, Carbonell [19] devised a queuing model in which different instruments compete for human attention. The model assumes that each time the human looks at an instrument, he or she postpones the observation of the other instruments, hence accepting the risk that another instrument exceeds a threshold. The optimal sampling strategy is then to sample, and bring back to zero, the instrument with the highest risk of not being observed. The momentary risk per instrument is defined in terms of the cost of exceeding a threshold value (cf., "importance" in [10]) and the probability that the instrument pointer will exceed the threshold, which accumulates as a function of the time since last sampling the instrument.

Carbonell's model was experimentally validated by Carbonell et al. [20] but has received little attention since then. Other models of visual sampling behavior were proposed by Sheridan [21] and Kvalseth [22]. However, their models have not been empirically evaluated using eye-tracking equipment.

Nowadays, ample research exists on the topic of visual attention. Borji and Itti [23] provided a review of more than 60 models of visual attention, most of which are bottom-up models (i.e., saliency models). In the last decades, several promising models that include elements of top-down (i.e., task-driven) attention have been developed. For example, Wolfe [24] presented a model that predicts reaction times in tasks where observers look for a target among distractor items. Similarly, Najemnik and Geisler [25] showed that humans can localize a target stimulus embedded in a cluttered environment in an efficient manner, by making eye movements that gain the most information about target location. Salvucci and Taatgen [26] presented a computational model that computes reaction times and performance for diverse multitasking conditions, whereas Sprague et al. [27] presented a model of visual behavior, which included a simulated humanoid that allocates gaze based on variables of reward and uncertainty. In an attempt to combine bottom up and top down cues in a comprehensive manner, Wickens et al. [28] introduced the saliency, effort, expectancy, and value (SEEV) model of visual behavior. This model defines the probability of sampling an instrument/area in terms of two bottom-up variables: 1) saliency (i.e., the extent to which the stimulus stands out with respect to its background) and 2) effort (i.e., the amount of eye/head movement required) and two top-down variables: 3) expectancy (equivalent to bandwidth, i.e., the perceived likelihood of change or event frequency) and 4) value (i.e., subjective importance of attending to events on the instrument, or the cost of missing them). The SEEV model has received widespread experimental support (e.g., [28]–[30]).

In summary, in the past decades, various models have been developed that describe how humans sample a dynamic system. Much of the current visual models of human monitoring seem to be conceptually based on the original studies by Senders [13], [15]

(e.g., [4], [31]). Indeed, the work of Senders is relatively influential in the human factors community, as demonstrated by the ample number of citations in Google Scholar (254 for Senders [13], and 144 for Senders [15]). Perhaps somewhat peculiarly, the work of Senders has hardly been replicated. An exception is Fleetwood [32], who performed three experiments using five participants each. In each experiment, participants viewed four dials as in [13] while eye movements were recorded with a head-mounted eye-tracker. The results of Fleetwood's experiments showed that participants' mean glance durations per dial were sensitive to various experimental manipulations, including bandwidth, threshold cross frequency, value (i.e., different points could be earned based on the correct detection of a pointer that had gone out of bounds), visual saliency (i.e., flashing dial), and the cost of making an observation (i.e., implemented as a time delay when a participant indicated that they would like to view a new dial). Although the design of Fleetwood is regarded as informative, it remains unclear to what extent Senders' results were replicated. Thus, considering that the experiments of Senders were conducted with only five participants and with limited hardware equipment, Senders' study deserves to be replicated and extended for the sake of better insight in human sampling behavior.

The aim of the present study was to replicate the experimental conditions of Senders [15], using a larger sample size (N = 86 versus 5) and an eye tracker camera with high temporal resolution (2000 versus 12 Hz). Additionally, whereas Senders' work was solely concerned with coarse dependent measures (i.e., glance rate and duration), we applied fully synchronized data recordings of 1) the six pointer signals, 2) participants' eye movements, and 3) participants' button press inputs. This allowed us to examine how participants distributed their attention across the dials as a function of the state of the dials. An additional factor is that we varied the eye-movement effort level (i.e., one of the parameters in the SEEV model) by changing the dial configuration from a low effort configuration (high bandwidth dials in the center of the bank of dials) to a high effort configuration (high bandwidth dials in the corners of the bank of dials).

In order to structure and interpret our results, we classified our findings according to three variables of the SEEV model: 1) bandwidth (expectancy), 2) effort, and 3) saliency. In our experiment, bandwidth and effort are independent (i.e., experimentally manipulated) variables, whereas saliency is referenced by the momentary state of pointers. Note that value is not an experimental variable in our study, nor in Senders' work: all six dials were assumed to have equal value (i.e., equal importance) for performing the task.

It is noted that part of the results of the same experiment is presented by Eisma et al. [33] in more concise form. Therein, Eisma et al. [33] were concerned with the broader methodological topic of assessing situation awareness through a correlation between

an aggregate visual sampling-to-environmental relational score in comparison to self-reported situation awareness using a freeze-probe questionnaire method. The present study is not concerned with these self-reports but only with objective measures: stimulus behavior, observer performance, and eye-movement data.

2. METHODS

A. Participants

Participants were 86 university students (21 female, 65 male) with a mean age of 23.44 years (SD = 1.52). The research was approved by the Human Research Ethics Committee of the TU Delft under the title "Update of Visual Sampling Behavior and Performance with Changing Information Bandwidth" (September 22, 2016). Written informed consent was obtained from all participants.

B. Apparatus and Procedures

The eye movements of the right eye were recorded at 2000 Hz using the SR Research EyeLink 1000 Plus eye tracker. Participants were asked to put their head in a head/chin rest support, which was adjusted to the participant's height to reduce neck and shoulder strain. The participants were asked to keep their head on the head support throughout the duration of the experiment to the best of their ability, allowing for breaks to counteract any discomfort if needed.

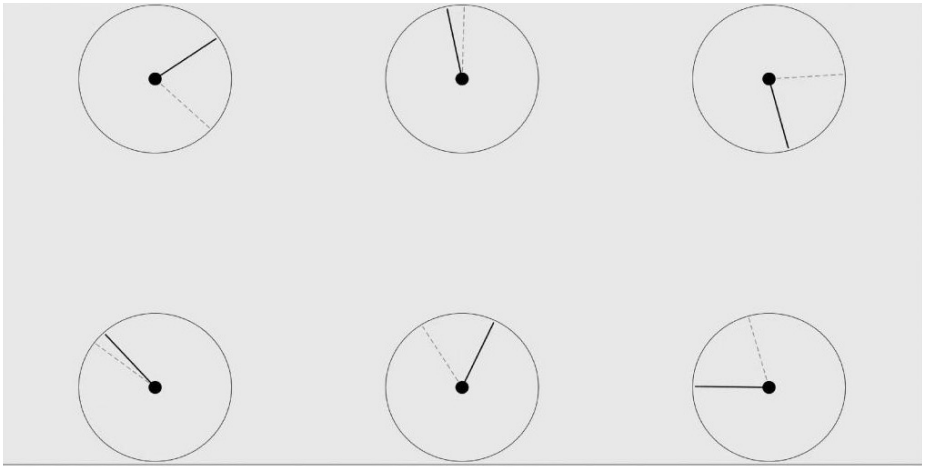


Figure 2. Screenshot of one of the seven videos. In each dial, the dashed line is the threshold and the solid line is the pointer.

The stimuli were presented on a 24 in BenQ XL2420T-B monitor with a resolution of 1920 × 1080 pixels (display area 531 × 298 mm^2), positioned approximately 95 cm in front of the participant and 35 cm behind the eye-tracking camera/IR light source. The stimulus display subtended approximately a 31° and 18° horizontal and vertical viewing angle, respectively.

First, the eye tracker was calibrated. Next, participants completed a 20 s familiarization trial, allowing them to get used to the experimental setup and task requirements. During this trial, a single dial was shown on the screen.

Next, participants viewed seven 90 s videos. Each video showed six circular dials with moving pointers. Each dial had a diameter of 316 pixels (visual angle ~5.3°), see Fig. 2. The centers of adjacent dials were 634 pixels (~10.5°) and 658 pixels (~10.9°) apart in horizontal and vertical direction, respectively, which is similar to [15] who reported that the dials in his experiments were separated by 12°. The dashed threshold line was a random angle that differed for each of the 42 dials (7 videos × 6 dials). In each of the seven videos, the pointer signals had a mean of 0° (i.e., the position of the threshold) and a standard deviation of 50.1°. The signal realization was different for each of the 42 dials. The MATLAB script that was used for creating the videos is provided in Appendix A.

The frame rate of the videos was 50 Hz, with a resolution of 1904 × 988 pixels. Each participant viewed the same seven videos but in a uniquely randomized order. Participants were instructed to press the spacebar when any of the pointers crossed the threshold from either direction.

After viewing each video, participants completed a brief questionnaire to probe their self-reported situation awareness, knowledge confidence, and experienced eye movement effort. The total time of experiment participation varied between 15 and 30 min.

C. Independent Variables

1) Bandwidth: The first independent variable was the bandwidth of the dials. The six pointers each had a different bandwidth: 0.03, 0.05, 0.12, 0.20, 0.32, and 0.48 Hz, as in [15]. More specifically, a signal was defined as a sum of 21–41 sinusoids, with random phase shifts and with predefined bandwidth (i.e., cutoff frequency) in agreement with Elkind [34] and Senders [15]. Naturally, the high bandwidth dials also moved more rapidly, with overall mean absolute angular pointer velocities of 6.2, 7.3, 13.6, 20.8, 35.4, and 43.5°/s for the 0.03, 0.05, 0.12, 0.20, 0.32, and 0.48 Hz dials, respectively. The videos are available as the Supplementary material (see Appendix B). The pointer movement of each of the six dials for one of the seven videos is shown in Fig. 3.

2) Effort: The second independent variable was the effort level. Each of the seven videos had a dial configuration that differed according to the predicted amount of eye-movement effort participants had to put in, in order to respond perfectly to each threshold crossing. The configurations were selected with the help of a computer simulation (see Appendix C for the script), in which a value of 1 was assigned to the distance between two adjacent dials (e.g., the diagonal distance

between two corner dials was determined as 5, see Appendix D for an overview of distances between pairs of dials). All dial configurations are shown in Table I.

Note that Senders similarly positioned the dials "in a quasi-random way in order to achieve as much counterbalancing as possible, since the theoretical model, which was to be tested, did not consider the factor of arrangement of signals of various frequencies" (see [17, p. 44]). However, Senders [15], [17] did not present the actual dial configurations.

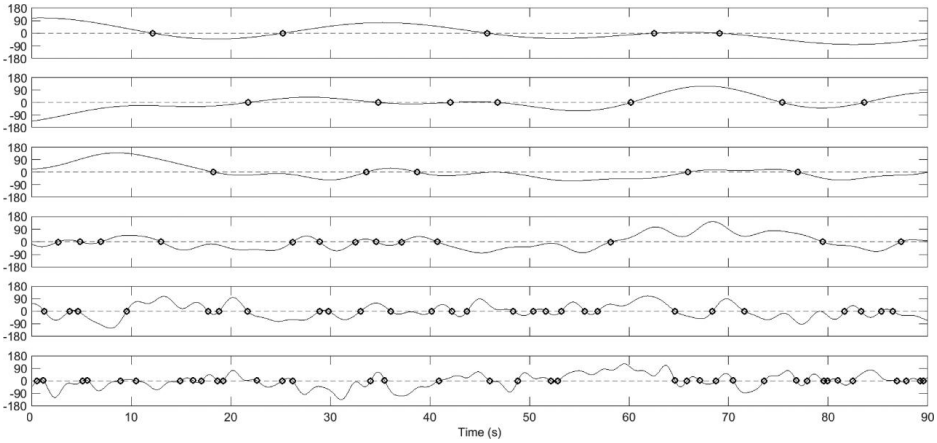


Figure 3. Pointer angle in degrees relative to the threshold (positive = clockwise with respect to the threshold, negative = counterclockwise with respect to the threshold) as a function of elapsed time in one of the videos. The six subplots are sorted on bandwidth (top = 0.03 Hz, bottom = 0.48 Hz). A circular marker indicates a threshold crossing.

Table I. Bandwidth (Hz) per dial position for each of the seven videos used in the experiment

Video effort level	Top left	Top middle	Top right	Bottom left	Bottom middle	Bottom right	Effort level
Level 1 (lowest effort)	0.12	0.48	0.05	0.20	0.32	0.03	3422
Level 2	0.20	0.48	0.03	0.32	0.05	0.12	3686
Level 3	0.03	0.12	0.20	0.32	0.48	0.05	3896
Level 4	0.32	0.12	0.05	0.48	0.03	0.20	4097
Level 5	0.48	0.05	0.03	0.12	0.20	0.32	4314
Level 6	0.12	0.32	0.05	0.20	0.03	0.48	4532
Level 7 (highest effort)	0.32	0.03	0.20	0.12	0.05	0.48	4969

The video numbers range from low effort (high bandwidth dials in the middle) to high effort (high bandwidth dials in the outer edges). The effort level is the estimated cumulative saccade distance if the participant were to sample perfectly for 1 h of observation.

D. Dependent Variables

1) Dependent Measures to Replicate Senders [15]: First, missing x and y coordinates during blinks were restored with linear interpolation. Furthermore, a median filter with a 100 ms interval was applied to the x and y gaze coordinates. Next, the following measures were calculated per participant, per dial, and per video:

    1) Glance rate (Hz), defined as the number of times per second that the participant fixated on a 420 × 420 pixel area of interest (AOI) surrounding the dial. Refixations on the same dial were not counted. By virtue of a fixation filter, only glances on dials were counted, not fly-throughs (e.g., the top middle dial was not counted when the participant performed a saccade from the top left to the top right dial). Gaze velocity data were calculated and filtered with a Savitzky–Golay filter with order 2 and a frame size of 20 ms (i.e., 41 samples at 2000 Hz, twice the minimum saccade duration of 10 ms, see [35]). We adopted a saccade velocity threshold of 2000 pixels/s (~33°/s). It has been reported that fixation durations in reading can be as short as 50–75 ms [36]. Considering that the present task involved rapid sampling and small visual angles, a minimum fixation duration of 40 ms was used, see also [35].

    2) Percent time on AOI (%), defined as the percentage of video time that the eye-gaze of the participant was within a specified dial AOI. This measure was calculated independently from the fixation filter and has also been referred to as the net dwell time percentage [37].

    3) Mean glance duration (s), defined as the net dwell time per dial in seconds divided by the number of glances on that dial.

These three preceding measures were compared to the corresponding measures reported in [15]. Note that Senders [15] used the terms 1) fixation frequency or sampling frequency, 2) percent time fixated, and 3) duration of fixation, for the three above-mentioned measures, respectively. However, for the sake of clarity, we adhered to modern terminology in line with standards [38].

Additional dependent measures were taken as follows.

2) Spacebar Press Performance: We calculated a performance score, defined as the percentage of threshold crossings for which the participant pressed the spacebar. In total, there were between 74 and 115 threshold crossings per video. Per crossing, a hit was counted if the participant pressed the spacebar within 0.5 s (i.e., between −0.5 and +0.5 s) of the moment of the crossing. Specifically, hits were determined using a forloop over the threshold crossings of a video in a chronological order. For each threshold crossing, the temporally closest spacebar was selected, and if

the absolute time difference between the moment of pressing the spacebar and the moment of the threshold crossing was smaller than 0.5 s, then that threshold crossing was labeled a hit, and the spacebar press was excluded from being assigned to subsequent threshold crossings. Accordingly, a spacebar press could not be assigned to more than one threshold crossing, and no more than one hit could be assigned to a threshold crossing.

3) Questionnaire Data: Per participant and per video, we calculated the experienced eye-movement effort. This measure was defined as the response to the question "How much eye-movement effort did you experience?," with response options from 1 (very low) to 10 (very high).

E. Analyses

In order to structure our findings, our analyses were categorized into 1) bandwidth (expectancy), 2) effort, and 3) saliency, which are the first three predictor variables of the SEEV model.

1) Bandwidth (Expectancy)—Replication of Senders [15]: First, we reported the overall glance rate, percent time on AOI, and mean glance duration as a function of bandwidth, in order to examine whether the results of Senders were replicated in our study. Similarities between our results and Senders' results were assessed by comparing the parameters of linear least squares fits between the bandwidth and the dependent measure.

   A periodic sampling model assumes that the human observer forms expectancies about the likelihood that a pointer will cross a threshold, or as pointed out by Senders [15]: "in order to make a rational allocation of visual attention to various signals, the observer must learn the bandwidths of those signals" (p. 86). To investigate whether participants exhibited learning (i.e., whether they formed expectancies) during the experiment, we assessed linear least squares fits between glance rate and bandwidth, per video presentation number in the chronological order. This allowed us to assess whether participants distributed their attention more akin to the Nyquist theorem as they gained experience at the sampling task.

2) Effort (Dial Configuration): Similar analyses were conducted for the different video effort levels. That is, we calculated linear fits between the mean glance rate and dial bandwidth, for each dial configuration condition shown in Table I. Additionally, the performance score and self-reported effort were computed per effort level, to see whether participants performed more poorly in the high effort videos than in the low effort videos.

3) Saliency (Pointer Angle, Pointer Velocity, Time to Crossing): Finally, we assessed whether participants were conditional samplers by calculating for each video frame

the percentage of participants who glanced at each dial and comparing this to the angle and velocity of the dial pointers at that video frame. Here, pointer angle (i.e., closeness to threshold) and pointer velocity are regarded as components of saliency, that is, the extent to which the current state of a pointer attracts visual attention. Note that for a single sine function, position and its derivative are directly related, as the derivative of a sine wave equals the same sine wave with a phase shift, and hence, in this case, it would be meaningless to analyze the effects of pointer angle and pointer velocity separately. However, in our experiment we used a multisine consisting of 40 aggregated sine waves, as a result of which pointer angle and pointer velocity were not directly related, except at its extreme values (i.e., when a pointer signal reaches its peak angle in a given video, the pointer has a velocity of zero by definition).

In addition to the pointer angle and pointer velocity, we assessed conditional sampling for the "time to crossing," defined as the momentary pointer angle divided by the signinverted pointer velocity (see [39] for a similar time to line crossing measure in car driving, and [40] for the notion that humans may be able to perceive time to crossing directly from the closure rate of the logtransformed angle between pointer and threshold). A positive time to crossing means that the dial is moving in the direction of the threshold, whereas a negative value means that the pointer is moving away from the threshold.

It is noted that vision researchers typically use the word saliency to refer to stimulus characteristics such as intensity contrast, flicker contrast, and motion contrast, devoid of task context [41]. Absolute pointer velocity is a saliency feature, but the pointer angle is not. That is, participants should interpret the pointer angle in relation to the task of pressing the spacebar when it crosses the threshold. Herein, we use the term saliency in a broad meaning, by defining it as the dial's momentary characteristics (as opposed to bandwidth, which is a timeinvariant property).

3. RESULTS

Data were lost for a few videos (1–3) from a total of three participants. However, because the majority of their data were still available and unaffected, these three participants were retained in the analysis.

A. Descriptive Statistics: Aggregate Gaze Results

Figure. 4 shows the aggregated distribution of all gaze coordinates on the monitor. It is apparent that not all six dials exhibited the same percent time on AOI. The percent time on AOI was the highest for the top middle dial (20.18%) position and the lowest for the top right dial (8.50%) position. These differences are consistent with a baseline tendency to look at the middle two dials, and can also be explained by the different

bandwidth configurations per dial (see Table I). For example, the top right dial position never happened to display a high bandwidth dial signal (0.32 or 0.48 Hz), which may explain why it was overall less sampled relative to the dial positions in the other corners. Further analysis (see Appendix D) showed that diagonal eye movements were rarer compared to horizontal and vertical ones, see [4] for a similar finding.

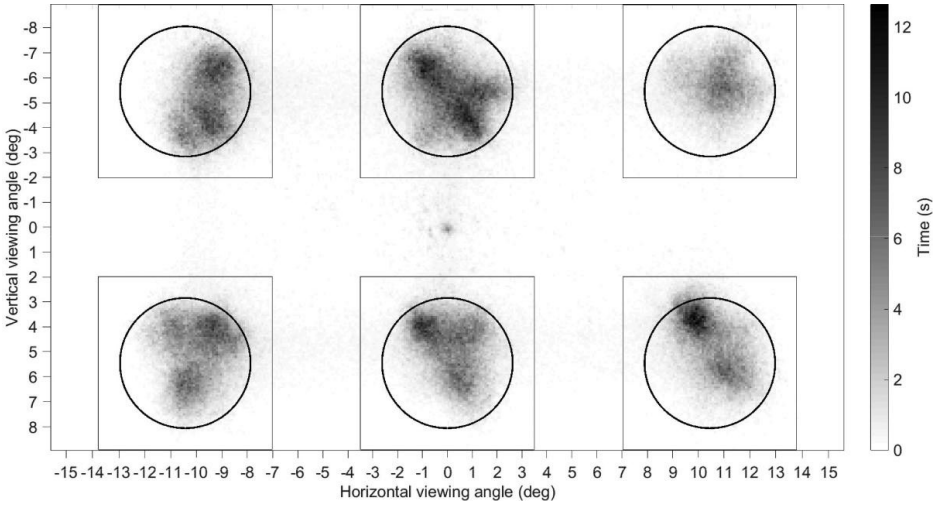


Figure 4. Distribution of gaze for all videos of all 86 participants aggregated (53,550 s of data). For the purposes of this visualization, the screen was divided into 5 × 5 pixel squares, and the darker the color, the more time was spent looking at that part of the screen as indicated by the vertical bar on the right. The circles represent the dials; the squares that surround the circles represent the areas of interest.

B. Descriptive Statistics: Aggregate Performance Results (Spacebar Presses)

Figure 5 shows the distribution of the time difference between the spacebar presses and the threshold crossings of the pointers. The 10th, 50th, and 90th percentiles of the time difference were −0.66, 0.06, and 0.68 s, respectively. Accordingly, our definition of performance, which incorporated a time margin from −0.5 to 0.5 s surrounding each threshold crossing, is regarded as reasonable in that it captured the majority of spacebar presses surrounding a threshold crossing while minimizing overlap between consecutive threshold crossings.

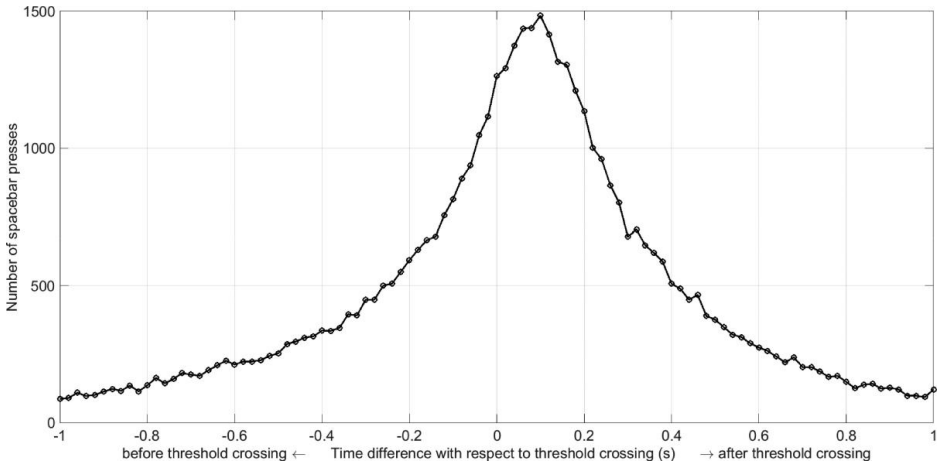


Figure 5. Time difference between threshold crossing and the spacebar press that occurred nearest in time, for each threshold crossing (N = 52,627). Results are presented in 0.02 s intervals.

Table II shows that participants slightly improved their spacebar-pressing performance score from 47.53% during the first video up to 51.17% in the last video, whereas the corresponding standard deviation among participants remained approximately constant. The effect of video presentation order was small but statistically significant according to a repeated measures ANOVA for the 83 participants without missing values, $F(6,492) = 5.37$, $p < 0.001$, $\eta_p^2 = 0.061$. Additionally, a higher effort configuration of the dials corresponds with a slightly lower performance score (see Table II). The effect of the video effort level was statistically significant as well, $F(6,492) = 14.14$, $p < 0.001$, $\eta_p^2 = 0.147$.

Table II. Performance scores as a function of video presentation order (i.e., learning effect in performance) and as a function of the video effort level (i.e., effect of dial configuration on performance)

Performance as a function of video presentation order		Performance as a function of video effort level	
Video order	Performance score (%) M (SD)	Video effort level	Performance score (%) M (SD)
First	47.53 (8.74)	Level 1 (lowest effort)	52.92 (8.39)
Second	48.49 (8.54)	Level 2	51.35 (8.66)
Third	49.40 (9.14)	Level 3	47.97 (8.22)
Fourth	48.42 (9.08)	Level 4	48.49 (8.43)
Fifth	49.82 (8.11)	Level 5	47.28 (8.27)
Sixth	51.40 (8.38)	Level 6	50.01 (8.68)
Seventh (last)	51.17 (8.63)	Level 7 (highest effort)	48.11 (9.23)

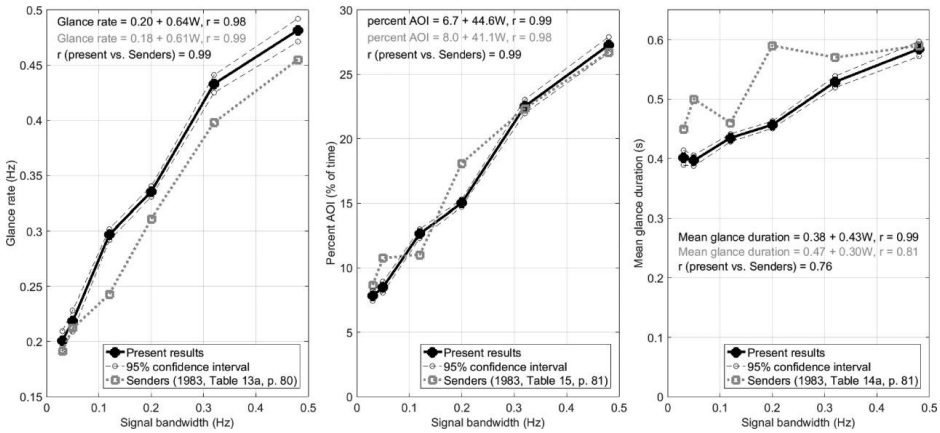


Figure 6. Glance rate, percentage of time on area of interest (AOI), and mean glance duration as a function of signal bandwidth of the dial. The dashed lines with open circles represent 95% confidence intervals around the mean, calculated according to Morey [41]. The grey dotted line with square markers corresponds to Senders' summary results in which he averaged the results of three similar experiments. The equations represent a least squares linear fit for our results (in black) and Senders' [15] results (in grey). Also shown is the Pearson correlation coefficient between our results and Senders' results.

We observed no interpretable relationship between dial position (i.e., top left, top middle, top right, bottom left, bottom middle, bottom right) and performance scores (see Appendix E). However, the lowest bandwidth dial (0.03 Hz) featured a lower performance score (30.96%) than the five higher bandwidth dials (46.55% and higher, see Appendix E). Further inspection revealed that the difficulty of the dials was highly idiosyncratic: among the 42 dials (7 videos × 6 dials) the performance score ranged between 16.00% (SD = 16.85%) for the top middle dial (0.03 Hz bandwidth) of the video with effort level 7, and 66.09% (SD = 20.20%) for the bottom right dial (0.12 Hz bandwidth) of the video with effort level 2.

C. Bandwidth (Expectancy)—Replication of Senders [15]

Figure 6 shows the glance rate, percent time on AOI, and mean glance duration, as a function of bandwidth, together with 95% confidence intervals for the means across the participants. Also shown are the results of Senders, which are based on a total of five participants. Because Senders used a small number of participants, no confidence intervals were calculated for his dataset. The results reveal a high correspondence between our results and Senders' [15] results (r = 0.99, 0.99, and 0.76 for the three respective measures).

In order to assess whether participants exhibited learning (i.e., whether they formed expectancies of bandwidth) from the first video to the seventh video, the glance rate as a function of bandwidth was assessed per video number. The results in Table III show that there is a slight learning effect, as the slope is shallowest 0.61 W for the first video

and steepest 0.68 W for the seventh video. Note that these changes in the parameters of the linear fits are overall small and that the parameters for all video presentation orders are in agreement with Senders who reported GR = 0.18 + 0.61 W, *r* = 0.99 (see Fig. 6).

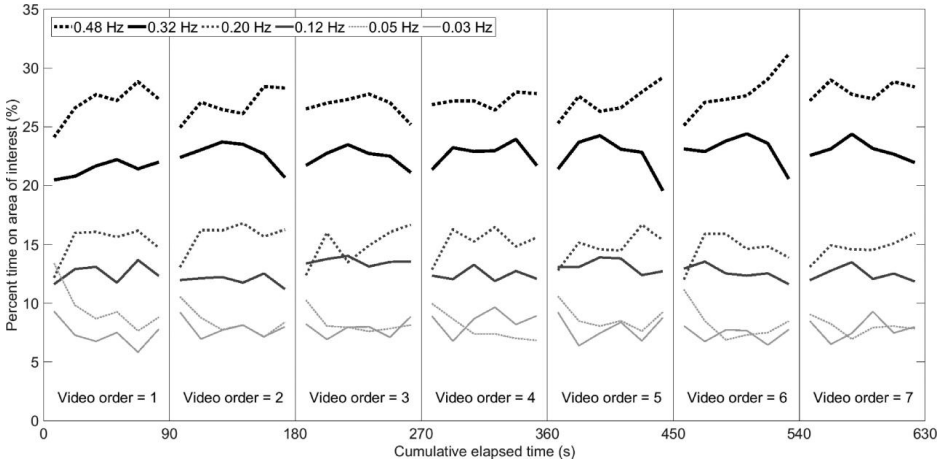


Figure 7. Percentage of time that participants had their eyes on a particular bandwidth dial as a function of the total elapsed video time. Each video lasted 90 s. The results are provided as averages per 15 s wide bin. These confidence intervals are depicted only for the lowest and highest bandwidth dials, in order to prevent clutter.

Table III. Linear fit for bandwidth (W) versus mean glance rate (GR) as a function of the chronological order of video presentation

Video presentation order	Linear fit and correlation coefficient (*r*)
First	GR = 0.21 + 0.61 W, *r* = 0.98
Second	GR = 0.20 + 0.64 W, *r* = 0.98
Third	GR = 0.20 + 0.63 W, *r* = 0.98
Fourth	GR = 0.20 + 0.65 W, *r* = 0.97
Fifth	GR = 0.21 + 0.62 W, *r* = 0.98
Sixth	GR = 0.20 + 0.66 W, *r* = 0.98
Seventh (last)	GR = 0.19 + 0.68 W, *r* = 0.98

Table IV. Linear fit for bandwidth (W) versus mean glance rate (GR), and self-reported effort, as a function of the video effort level (see Table I for definition)

Video effort level	Linear fit and correlation coefficient (*r*)	Self-reported effort M *(*SD)
Level 1 (lowest effort)	GR = 0.11 + 1.00 W, *r* = 0.97	6.89 (1.61)
Level 2	GR = 0.16 + 0.84 W, *r* = 0.95	6.84 (1.59)
Level 3	GR = 0.21 + 0.65 W, *r* = 0.86	7.01 (1.62)
Level 4	GR = 0.23 + 0.50 W, *r* = 0.82	6.91 (1.86)
Level 5	GR = 0.26 + 0.38 W, *r* = 0.71	7.16 (1.49)
Level 6	GR = 0.18 + 0.66 W, *r* = 0.84	7.24 (1.62)
Level 7 (highest effort)	GR = 0.23 + 0.44 W, *r* = 0.96	7.19 (1.54)

Figure 7 presents a further illustration of the learning effect within a particular video. It can be seen that the percent time on the dials with different bandwidths can already be differentiated from the beginning (i.e., in the first 15 s of each 90 s). There also appears to be a slight periodicity for each of the seven videos as it seems to take about 30 s for AOI percentages to settle in (e.g., the 0.20 Hz dial appears to be relatively undersampled in the first 15 s of each video). A paired t-test between the first 15 s and last 15 s indicated the following: t(82) = 1.42, 6.03, −0.36, −3.30, −0.90, −2.64 ($p = 0.160$, < 0.001, 0.718, 0.001, 0.370, 0.010, Cohen's $d_z = 0.16$, 0.66, −0.04, −0.36, −0.10, −0.29) for the 0.03, 0.05, 0.12, 0.20, 0.32, and 0.48 Hz dials, respectively. In other words, the low bandwidth dials tended to be sampled less while the high bandwidth dials tend to be sampled more in the last 15 s as compared to the first 15 s. Thus, slight learning/habituation effects are distinguishable: sampling becomes more distributed with experience, which is in line with the increasing slope from 0.61 W to 0.68 W shown in Table III.

D. Effort (Dial Configuration)

The results in Table IV show that the slope of the regression line between bandwidth and glance rate was steepest (1.00 W) for the lowest effort configuration and considerably shallower (0.44 W) for the highest effort configuration. To illustrate, in the lowest effort configuration, participants had a glance rate of 0.128 and 0.554 Hz to the low and high bandwidth dial, respectively. In the highest effort configuration, this was 0.252 and 0.429 Hz, respectively. In other words, when the effort was lower, the effect of bandwidth on distributed sampling was higher. Thus, when the high bandwidth dials were placed in the middle (e.g., video effort level 1) instead of at the outer edges (e.g., video effort level 7), participants behaved more in accordance with the Nyquist theorem (i.e., a slope that is closer to the theoretically predicted slope of 1 W or 2 W, depending on whether or not sampling of the velocity is taken into consideration).

Conversely, when the high bandwidth dials were placed at the outer edges, participants relatively rarely sampled these high bandwidth dials while relatively often sampling the low bandwidth dials in the middle, in line with the notion that effort inhibits sampling.

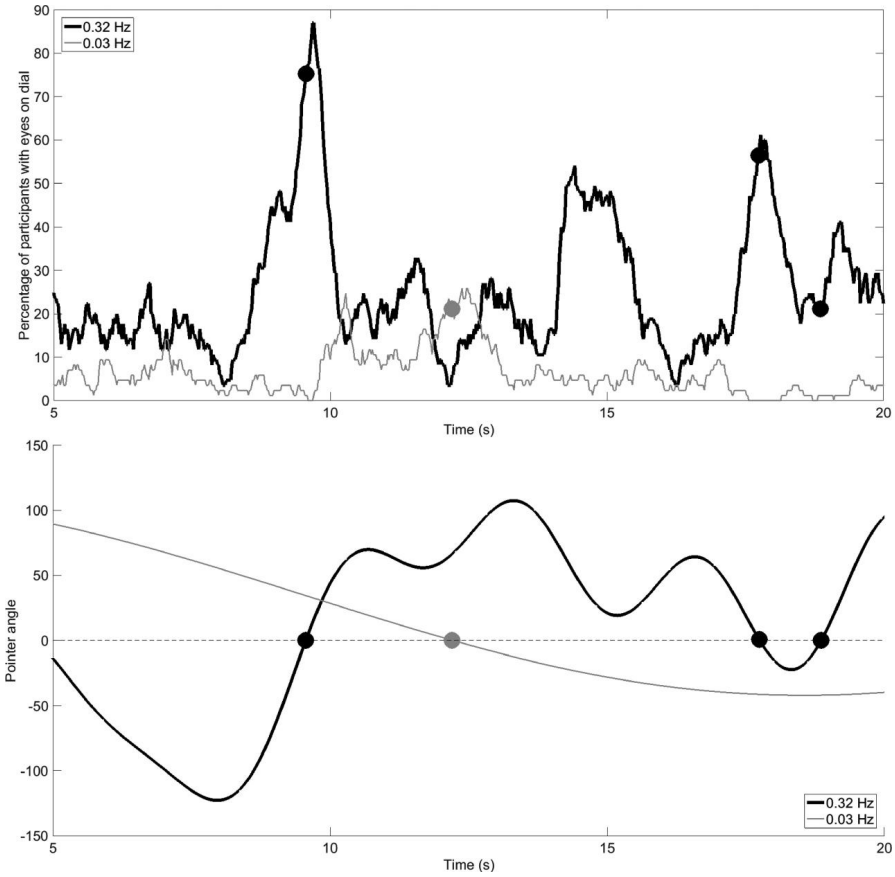


Figure 8. Percentage of participants (N = 85) with their eyes on a dial (top) and state of the dial relative to the threshold (bottom) for a representative 15 s segment of the first of seven videos. The circular markers are depicted at the moments of the threshold crossings.

Table IV also shows that objective effort had a small effect on subjective effort; this effect was significant according to a repeated measures ANOVA, $F(6,498) = 2.78$, $p = 0.012$, $\eta_p^2 = 0.032$.

E. Saliency (Pointer Angle, Pointer Velocity, Time to Crossing)

An initial exploration confirmed that participants' sampling behavior was indeed not only dependent on bandwidth and effort, but also highly time-varying. Figure 8 shows the percentage of participants who gazed at two specific dials for a random 15 s segment of one of the seven videos. Once again, it is evident that participants looked more at a high bandwidth dial than at a low bandwidth dial. Closer inspection shows that

participants were more likely to gaze at a particular dial (see peaks in the upper graph) when the pointer angle was near the threshold (see bottom graph having the same time axis as the upper graph). Furthermore, peaks in the upper graph appear to occur when a pointer angle has a high gradient, that is, when the pointer was moving rapidly.

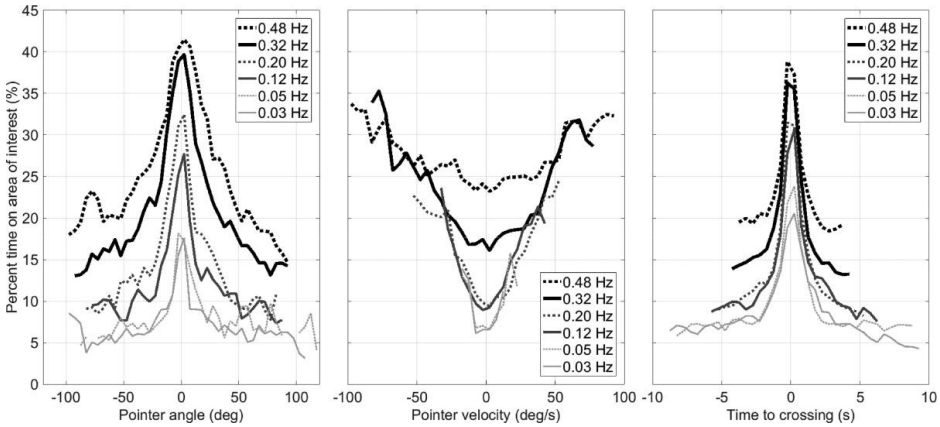


Figure 9. Relationship between pointer angle in 5° increments (left), pointer velocity in 5°/s increments (middle), and time to crossing in 0.5 s increments (right) versus percent time on area of interest. The results in this figure were based on all videos of all participants. Only data points for which at least 5 s of video data were available are shown.

The relationship between the participants' viewing behavior and the state of the dials is further illustrated in Fig. 9. The left panel shows the proportion of participants sampling a specific dial as a function of the pointer angle with respect to the threshold. It is clear that participants were considerably more likely to gaze at a dial when the dial was close to the threshold. When the dial was near the threshold at 0°, the probability of sampling the dial was about 2-3 times as high as compared to when the dial was at an angle of 45° away from the threshold. In the middle and right panel of Fig. 9, this effect can also be seen for the angular velocity of the pointer and time to crossing, respectively. Note that there appears to be no asymmetry: high-velocity pointers attract attention regardless of whether the pointer is moving toward or away from the threshold. Figure 10 further illustrates that a combination of low pointer angle and high pointer velocity is an attractor of attention. Also, it is notable that the low bandwidth dials never reach a high velocity in the first place.

In sum, participants do not behave in accordance with a periodic sampling model. Rather, participants sample conditionally: the closer the pointer to the threshold, and/or the faster it moves especially toward that threshold, the more the participants gazed at that specific dial.

4. DISCUSSION

A. Bandwidth (Expectancy)—Replication of Senders [15]

The aim of this research was to replicate Senders' [15] study of visual sampling, using high-end eye-tracking equipment, and a larger number of participants. The results of our experiment showed that the glance rate, the percent time on AOI, and the mean glance duration increase as the bandwidth increases, in close similarity to what was found by Senders (i.e., highly similar slopes and intercepts, and strong correlations between our results, r = 0.99, 0.99, and 0.76). In his work, Senders [15] noted that the high bandwidth dials do have a longer duration of observation, but he also expressed considerable uncertainty about this effect. Our results confirm for the first time a linear relationship between bandwidth and mean glance duration (r = 0.99, see Fig. 6). Presumably, we obtained a stronger correlation with bandwidth than the correlation obtained by Senders (r = 0.81) because Senders used only five participants in his experiment, hence giving rise to a considerable sampling error. Furthermore, Senders used manual coding of film recordings in lieu of eye tracking. These film images were recorded at 12 Hz, which means that the temporal resolution of his method was at best 0.083 s, or perhaps only 0.167 s if considering that at least two frames are needed to ascertain whether the eyes of the participant have actually landed on the dial. In comparison, we used an eye tracker with 2000 Hz resolution, combined with a fully automated data analysis procedure, which is insensitive to manual coding errors.

Senders argued that people need extensive training in order to learn the statistical characteristics (i.e., bandwidths) of the pointers: "The theory and the attendant models, therefore, apply only to demanding tasks performed by experienced and skilled human beings. No novices need apply." (see [15, p. 21], and also [16]). The participants in our experiment were only allowed to familiarize for 20 s with a single dial (as an example), yet our results show considerable similarities with Senders' results. This suggests that participants do not need to learn the statistical characteristics of the signal, but predominantly rely on the momentary state of the dial in order to perform the sampling task.

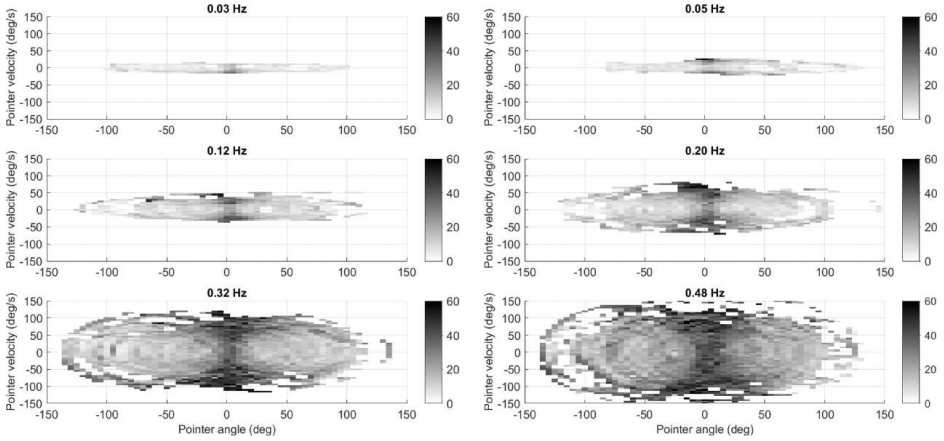


Figure 10. Percent time on area of interest (as indicated by the vertical bar next to each figure) as a function of pointer position and pointer velocity. The present figure shows the probability that participants sampled a dial for a combination of pointer position and pointer velocity. Pointer angle and pointer velocity were divided into 5° and 5°/s increments, respectively.

In our experiment, participants did show learning, as they distributed their attention more according to bandwidth in later sessions (see Table III; Fig. 7). However, this learning effect was minor compared to the effects of bandwidth and effort (dial configuration; Table IV). Furthermore, a comparison between the first 15 s and last 15 s (see Fig. 7) showed that the learning effect was bandwidth-specific (e.g., the 0.05 Hz dial showed a larger learning effect than the 0.03 Hz dial), which may be due to interactions with the dial configuration or the specific properties of the pointer signals.

B. Effort (Dial Configuration)

The information access effort in the context of this experiment can be defined as the amount of eye-movement required in order to detect the events (i.e., the threshold crossings). According to Wickens [42], people tend to minimize effort during the task, hence try "to avoid longer scans or other information access travels when shorter ones can be made" (p. 54). Our results provide support to the notion that objective effort inhibits sampling: people are more likely to gaze at high bandwidth dials when these are placed centrally and generally sample less in accordance to bandwidth as required eye-movement distances grow (see Table IV for the corresponding linear regression results).

C. Saliency (Pointer Angle, Pointer Velocity, Time to Crossing)

Senders [13] originally modeled human sampling behavior by assuming that the human acts as "a random sampling device constrained only by the base probabilities of each of the things sampled" (p. 5, emphasis added). However, for a six dial configuration, the slope of glance rate versus bandwidth is considerably shallower than the theoretically predicted 2.0. Both Senders [15] and ourselves found a slope of about 0.60 combined with an intercept of about 0.20 (see Fig. 6), indicating that participants undersampled

the higher bandwidth dials and oversampled the lower bandwidth dials relative to an assumed perfectly matched bandwidth dependent sampling behavior (i.e., slope of 2.0 and intercept of 0).

According to Senders [15], this shallow slope may be attributed to 1) mental overload, 2) the fact that participants exhibit forgetting of the pointer state since the last glance on the dial, and also 3) the fact that participants may sample not only the pointer angle but also the pointer velocity (i.e., additional information in the task such as stimulus saliency). The latter explanation is in agreement with the extended sampling theorem [18], which postulates that the slope of an observer is 1 W instead of 2 W when the observer extracts both momentary velocity and momentary position. Our results in Figs. 8–10 indicate that participants were more likely to glance toward a dial when the velocity of that dial's pointer was higher.

In sum, the periodic sampling model is contentious because the probability of sampling is strongly dependent on how close the dial is to the threshold and how rapidly the pointer is moving. It is striking that there is a strong U-shape for low bandwidth dials in particular: for relatively high pointer velocities (around −40 or 40°/s), the dwell percentages for low bandwidth dials are about equal to the dwell percentages for high bandwidth dials (see Fig. 9, middle). In other words, the pointer velocity can be a strong attention attractor even when the dial bandwidth is relatively low.

We argue that participants were able to detect whether something is happening quickly in the periphery, resulting in a state of uncertainty, which in turn attracts attention. The notion of motion being an attention attractor corresponds to the saliency cue in Wickens' SEEV model of visual sampling and many other types of bottom-up visual attention models [23]. Previous research shows that humans can perform a control task such as car driving [43] or pitch tracking [44] using peripheral vision.

Senders et al. [45] specifically examined whether it is possible to read a dial using peripheral vision, and one of their conclusions was that "an observer can discriminate among settings which differ by 45° almost perfectly even when the instrument is played as much as 40° from the line of sight" (p. 436). In a pilot experiment using larger dials and red thresholds that were placed upright, we found that a participant could complete the spacebar-pressing task satisfactorily while looking only at the center of the screen. Although the present layout (i.e., smaller dials, dashed threshold at various angles; see Fig. 2) is more difficult to perform using just peripheral vision, it is likely that participants are still able to extract some information from their periphery. In sum, we argue that participants do not have to rely only on the learned bandwidth of a signal to determine where to look (as predicted by the periodic model); rather, they detect in their periphery salient aspects of whether a dial's threshold is likely to be crossed (e.g.,

from a pointers' velocity, threshold proximity, or closure rate), which in turn attracts their foveal attention toward that dial.

D. Conclusion and Recommendations

Collectively, our results offer a more fine-grained picture of human visual sampling than that of periodic signal reconstruction according to the (extended) Nyquist-Shannon sampling theorem. In particular, our results indicate that even for a simple paradigm of six moving dials, human visual sampling should not be explained in terms only of bandwidth (expectancy) but also by effort and saliency, as used in the SEEV model [28]. In conclusion

1) the results of Senders [15] have been replicated using high-end eye-tracking equipment,
2) humans do not behave as periodic samplers, but as conditional samplers instead, and
3) the conditions upon which humans sample include aspects of both "saliency" and "effort" in addition to "expectancy" (i.e., base bandwidths) when considering visual sampling behavior in goal-directed task environment of a certain performance value.

Future research could be directed toward resolving some uncertainties in the present findings. First, although there is a close correspondence between the results obtained by Senders [15] and the results presented herein, it cannot be established what exactly caused the similarity of results. We closely reproduced Senders' signal composition and task instructions, but there are also some evident differences between our experiments. That is, we used a computer screen, whereas Senders used micro-ammeters, and we used a single randomly oriented threshold per dial (see Fig. 2), whereas Senders used a fixed threshold at about 56° on either side for all six dials.

Additionally, Senders provided participants with more than 10 h of training, whereas we provided essentially no training. In future research, these effects could be studied independently in more detail. It may be worthwhile to investigate how with elongated practice the components of the SEEV model come into play in a different manner. With extended exposure, the impact of saliency might be expected to slightly decrease due to habitation, while the impact of effort may also slightly decrease, as practice may encourage the development of more efficient motor/behavioral patterns to some limit.

The present study suggests that participants sample conditionally rather than periodically, and that peripheral saliency is an important attractor of attention. Future research could examine whether participants still learn some of the signal properties so that they can direct attention as a function of bandwidth even if peripheral vision is unavailable. For example, future research may use occlusion techniques [cf., 46] or a gaze-contingency paradigm.

We also recommend research into different dimensions of effort, such as eye movement effort, head movement effort, and cognitive effort. For example, it would be interesting to examine what happens if head movement is not restricted by a head support. Here, it might be expected that participants will orient their head toward the high bandwidth stimuli, and accordingly mitigate their required sampling effort. Future research may also investigate the interaction between physical effort and cognitive effort. For example, research in natural tasks has found that people tend to rely more on memory if the task requires more head movement [47].

It should be noted that the SEEV model served as a qualitative structure (i.e., bandwidth, effort, and saliency) for presenting our results. Our aim was not to compare different sampling models, and it, therefore, remains to be investigated whether the SEEV model yields a better fit to the data than queuing models of visual sampling and other models that use uncertainty and reward/cost of having (in)accurate state information [20], [48]. An inelegance of the SEEV model is that saliency is causally related to bandwidth, because higher bandwidth dials move faster and are, therefore, overall more salient. Furthermore, it is debatable whether bandwidth and value are orthogonal variables, as participants may believe that faster moving (higher bandwidth) dials are also the more important (higher value) dials. We also note that it is difficult to compare the relative contributions (e.g., in terms of variance explained) of bandwidth, effort, and saliency, because bandwidth differs between dials, effort differs between videos, and saliency (e.g., whether a pointer moves fast or not) differs per dial as a function of elapsed time. The criterion with which models can be compared also deserves further examination.

Our results may have various implications for the design of human machine interfaces for the supervisory control of automated processes. In particular, the results make clear how different stimuli conditions and dial configurations compete for attention. For example, we found that it is less likely that an observer gazes to a particular instrument when this instrument requires transition effort, which reinforces the notion that instruments should be within visual reach. It is recommended to investigate whether the present results generalize to more complex tasks, such as supervisory control on-board the cockpit of an aircraft or an automated car, where the operator must not only monitor the instruments but also visually sample competing stimuli of the external environment.

The present work opens up opportunities to provide human operators with real-time feedback when it is predicted that visual and so subsequent task performance may degrade. For example, when a situation is dangerous yet signal indicators do not evidently reveal a danger (e.g., low dial velocity, operator habituated to low signal bandwidth), then that signal could be augmented by temporarily enhancing saliency, to in turn improve task performance.

REFERENCES

[1] P. Hancock, “Automation: How much is too much?” Ergonomics, vol. 57, no. 3, pp. 449–454, Dec. 2013.

[2] T. B. Sheridan, Humans and Automation: Systems Design and Research Issues. Santa Monica/New York, NY, USA: Hum. Factors Ergonom. Soc./Wiley, 2002.

[3] T. B. Sheridan, “Human centered automation: oxymoron or common sense?” in Proc. Intell. Syst. 21st Century IEEE Int. Conf. Syst., Man Cybern., 1995, pp. 823–828.

[4] M. Donk, “Human monitoring behavior in a multiple-instrument setting: Independent sampling, sequential sampling or arrangement-dependent sampling,” Acta Psychol., vol. 86, no. 1, pp. 31–55, Jun. 1994.

[5] K. S. Steelman, J. S. McCarley, and C. D. Wickens, “Theory-based models of attention in visual workspaces,” Int. J. Hum. Comput. Interact., vol. 33, no. 1, pp. 35–43, Sept. 2016.

[6] C. A. Gainer and R. W. Obermayer, “Pilot eye fixations while flying selected maneuvers using two instrument panels,” Hum. Factors, vol. 6, no. 5, pp. 485–500, Oct. 1964.

[7] A. Haslbeck and B. Zhang, “I spy with my little eye: Analysis of airline pilots’ gaze patterns in a manual instrument flight scenario,” submitted for publication.

[8] M. M. Heiligers, T. V. Holten, and M. Mulder, “Predicting pilot task demand load during final approach,” Int. J. Aviation Psychol., vol. 19, no. 4, pp. 391–416, Sept. 2009.

[9] A. A. Spady, “Airline pilot scan patterns during simulated ILS approaches,” NASA Tech. Paper 1250, Hampton, VA, USA, 1978. [Online]. Available: http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa. gov/19790011511.pdf.

[10] P. M. Fitts, R. E. Jones, and J. L. Milton, “Eye movements of aircraft pilots during instrument-landing approaches,” Aeronaut. Eng. Rev., vol. 9, no. 2, pp. 56–66, 1950.

[11] S. J. Landry, “Human-computer interaction in aerospace,” in The Human-Computer Interaction Handbook, A. Sears and J. A. Jacko, Eds. New York, NY, USA: Taylor & Francis, 2008, pp. 721–740.

[12] J. J. Seeberger and W. W. Wierwille, “Estimating the amount of eye movement data required for panel design and instrument placement,” Hum. Factors, vol. 18, no. 3, pp. 281–292, Jun. 1976.

[13] J. W. Senders, “The human operator as a monitor and controller of multide-gree of freedom systems,” IEEE Trans. Hum. Factors Electron., vol. HFE-5, no. 1, pp. 2–5, Sep. 1964.

[14] C. Shannon, “Communication in the presence of noise,” Proc. IRE, vol. 37, no. 1, pp. 10–21, Jan. 1949.

[15] J. W. Senders, “Visual sampling processes,” Ph.D. dissertation, Katholieke Hogeschool Tilburg, the Netherlands, 1983.

[16] N. Moray, “Monitoring behavior and supervisory control,” in Handbook of Perception and Human Performance, K. Boff, L. Kaufman, and J. Thomas, Eds. New York, NY, USA: Wiley, 1986, pp. 40:1–40:55.

[17] J. W. Senders, J. I. Elkind, M. C. Grignetti, and R. Smallwood, “An investigation of the visual sampling behaviour of human observers,” Rep. no. NASA CR-434 , Nat. Aeronaut. Space Admin., Washington, DC, USA, 1966.

[18] L. Fogel, “A note on the sampling theorem,” IEEE Trans. Inf. Theory, vol. 1, no. 1, pp. 47–48, Mar. 1955.

[19] J. Carbonell, “A queueing model of many-instrument visual sampling,” IEEE Trans. Hum. Factors Electron., vol. HFE-7, no. 4, pp. 157–164, Dec. 1966.

[20] J. Carbonell, J. Ward, and J. Senders, “A queueing model of visual sampling: Experimental validation,” IEEE Trans. Man Mach. Syst., vol. 9, no. 3, pp. 82–87, Sept. 1968.

[21] T. Sheridan, “On how often the supervisor should sample,” IEEE Trans. Syst. Sci. Cybern., vol. 6, no. 2, pp. 140–145, Apr. 1970.

[22] T. O. Kvalseth,“Human information processing in visual sampling,” Ergonomics, vol. 21, no. 6, pp. 439–454, Jun. 1978.

[23] A. Borji and L. Itti, “State-of-the-art in visual attention modeling,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 1, pp. 185–207, Jan. 2013.

[24] J. M. Wolfe, “Guided search 4.0: Current progress with a model of visual search,” in Integrated Models of Cognitive Systems, W. D. Gray, Ed. London, U.K.: Oxford Univ. Press, 2007

[25] J. Najemnik and W. S. Geisler, “Optimal eye movement strategies in visual search,” Nature, vol. 434, pp. 387–391, Mar. 2005.

[26] D. D. Salvucci and N. A. Taatgen, “Threaded cognition: An integrated theory of concurrent multitasking,” Psychol. Rev., vol. 115, no. 1, pp. 101-130, 2008

[27] N. Sprague, D. Ballard, and A. Robinson, “Modeling embodied visual behaviors,” ACM Trans. Appl. Perception, vol. 4, no. 2, pp. 1–23, Jul. 2007.

[28] C. D. Wickens, J. Goh, J. Helleberg, W. J. Horrey, and D. A. Talleur, “Attentional models of multitask pilot performance using advanced display technology,” Hum. Factors, vol. 45, no. 3, pp. 360–380, Sept. 2003.

[29] W. J. Horrey, C. D. Wickens, and K. P. Consalus, “Modeling drivers’ visual attention allocation while interacting with in-vehicle technologies,” J. Exp. Psychol., Appl., vol. 12, no. 2, pp. 67–78, Jun. 2006.

[30] C. Wickens, J. McCarley, and K. Steelman-Allen, “NT-SEEV: A model of attention capture and noticing on the Flight Deck,” Proc. Hum. Factors Ergonom. Soc. Annu. Meeting, vol. 53, no. 12, pp. 769–773, Jan. 2009.

[31] M. D. Byrne and R. W. Pew, “A history and primer of human performance modeling,” Rev. Hum. Factors Ergonom., vol. 5, no. 1, pp. 225–263, Jan. 2009.

[32] M. D. Fleetwood, “Refining theoretical models of visual sampling in supervisory control tasks: Examining the influence of alarm frequency, effort, value, and salience,” Ph.D. dissertation, Rice Univ., Houston, TX, USA, 2005.

[33] De Winter, J. C. F., Eisma, Y. B., Cabrall, C. D. D., Hancock, P. A., & Stanton, N. A. (2019). Situation awareness based on eye movements in relation to the task environment. *Cognition, Technology and Work,* 21, 99–111.

[34] J. I. Elkind, “Characteristics of simple manual control systems,” Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, MA, USA, 1956.

[35] M. Nystrom and K. Holmqvist, “An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data,” Behav. Res. Methods, vol. 42, no. 1, pp. 188–204, Feb. 2010.

[36] K. Rayner, “Eye movements and attention in reading, scene perception, and visual search,” Quart. J. Exp. Psychol., vol. 62, no. 8, pp. 1457–1506, Jun. 2009.

[37] SensoMotoric Instruments GmbH, BeGaze Manual. Teltow, Germany: SMI, 2014.

[38] Road Vehicles—Measurement of Driver Visual Behaviour With Respect to Transport Information and Control Systems—Part 1: Definitions and Parameters, ISO Standard 15007-1, 2014.

[39] H. Godthelp, P. Milgram, and G. J. Blaauw, “The development of a time-related measure to describe driving strategy,” Hum. Factors, vol. 26, no. 3, pp. 257–268, Jun. 1984.

[40] R. Bootsma and R. R. D. Oudejans, “Visual information about time-to-collision between two objects,” J. Exp. Psychol., Hum. Perception Per-form., vol. 19, pp. 1041–1052, 1993.

[41] R. D. Morey, “Confidence intervals from normalized data: A correction to Cousineau (2005),” Tutorial Quant. Methods Psychol., vol. 4, no. 2, pp. 61–64, 2008.

[42] C. D. Wickens, “Visual attention control, scanning, and information sampling,” in Applied Attention Theory, C. D. Wickens and J. S. McCarley, Eds. Boca Raton, FL, USA: CRC Press, 2008, pp. 41–61.

[43] H. Summala, T. Nieminen, and M. Punto, “Maintaining lane position with peripheral vision during in-vehicle tasks,” Hum. Factors, vol. 38, no. 3, pp. 442–451, Sept. 1996.

[44] A. Popovici and P. M. T. Zaal, “Effects of retinal eccentricity on human manual control,” in Proc. IMAGE Conf., 2017.

[45] J. W. Senders, I. B. Webb, and C. A. Baker, “The peripheral viewing of dials,” J. Appl. Psychol., vol. 39, no. 6, pp. 433–436, Dec. 1955.

[46] J. W. Senders, A. B. Kristofferson, W. H. Levison, C. W. Dietrich, and J.L. Ward, “The attentional demand of automobile driving,” Highway Res. Rec., vol. 195, pp. 15–33, 1967.

[47] D. H. Ballard, M. M. Hayhoe, and J. B. Pelz, “Memory representations in natural tasks,” Memory, vol. 7, no. 1, pp. 66–80, Dec. 2007.

[48] N. Sprague and D. Ballard, “Eye movements for reward maximization,” in Proc. Adv. Neural Inf. Process. Syst., 2003, pp. 1467–1474.

SUPPLEMENTARY MATERIALS

Appendix A: MATLAB script for creating videos

```
% Tested in Matlab 2015a
function create_videos(varargin)
clear variables;clc;close all
Dial_config=[3 6 2 4 5 1 % this matrix designates the bandwidth of each dial (top left, top
middle, top right, bottom left, bottom middle, bottom right), for video 1 to 7
    4 6 1 5 2 3
    1 3 4 5 6 2
    5 3 2 6 1 4
    6 2 1 3 4 5
    3 5 2 4 1 6
    5 1 4 3 2 6];
Rotation_matrix =  [3.38250527945727         5.4581698091647         0.914448383175152
3.22484196780935          5.67193259837401          2.44880133403221;
    6.25889901548155          0.530526063913437       0.854943968847379        2.52463433397692
5.93627298887054          1.51859113652713;
    0.491191333642999         2.51190846688352        5.46192402668991         0.477312801066939
3.08419005361748          2.53785485855062;
    2.78142960046077          1.63281389696207        3.64239134583996         1.50743765096339
3.07406498909385          0.606041655349003;
    0.670119118568649         5.02697851968287        3.45487354118084         0.774835719453349
2.12195363373904          0.829212653044237;
    6.04378388863272          2.71065302203305        0.910777858404502        1.15552671321197
5.65520510348193          5.91907843058038;
    0.0291176889893496        5.72176758533801        5.35975258543718         1.50766618367835
2.32004594985789          6.0075704948588];
t_end = 90;
cutoff = [.03; .05; .12; .20; .32; .48];
for kk=1:size(Dial_config,1) % loop over 7 videos
    [alldata,t] = signal_for_dial(t_end,kk); % create signals
    d=clock;save(strcat(['time', int2str(t_end), 'data_video', int2str(kk), '_config',
int2str(kk) '_CreatedAt_' [num2str(d(1)) '_' num2str(d(2)) '_' num2str(d(3)) '_' num2str(d(4))
'_' num2str(d(5))]]) , 'alldata');
    Xdata_begin = NaN(length(t),length(cutoff));
    Ydata_begin = NaN(length(t),length(cutoff));
    Xdata_end = NaN(length(t),length(cutoff));
    Ydata_end = NaN(length(t),length(cutoff));

    for k=1:length(cutoff) % Fill the Xdata and Ydata vectors
        Xdata_begin(:,k)=.079*sin(alldata(:,k)+Rotation_matrix(kk,k));
        Ydata_begin(:,k)=.079*cos(alldata(:,k)+Rotation_matrix(kk,k));
        Xdata_end(:,k)=.95*sin(alldata(:,k)+Rotation_matrix(kk,k));
        Ydata_end(:,k)=.95*cos(alldata(:,k)+Rotation_matrix(kk,k));
    end

    % make figure and animation
    v = VideoWriter(strcat(['Length', int2str(t_end),'_VideoNr', int2str(kk), '_Config',
int2str(kk) '_CreatedAt_' [num2str(d(1)) '_' num2str(d(2)) '_' num2str(d(3)) '_' num2str(d(4))
'_' num2str(d(5))]]), 'Mpeg-4'); %#ok<TNMLP>
    v.FrameRate = 50;
    open(v);

    close all
    figure
    set(gcf,'units','pixels','position', [0 0 1920 1080],'outerposition', [0 0 1920 1080])
    NF=length(t);
    for i = 1:NF+1
        for dial_nr=1:6
            h(dial_nr)=subplot(2,3,dial_nr);
            dial_frame(kk ,Dial_config(kk,dial_nr),Rotation_matrix);
            if i<=NF
                line1 = plot([Xdata_begin(i,Dial_config(kk,dial_nr)), Xdata_end(i,Dial_
config(kk,dial_nr))], [Ydata_begin(i,Dial_config(kk,dial_nr)), Ydata_end(i,Dial_config(kk,dial_
nr))]);
                line1.LineWidth = 2;
                line1.Color = [0 0 0];
            end
            hold off
        end
        set(h(1),'position',[0      0.6666 0.3333 0.3333])
        set(h(2),'position',[0.3333 0.6666 0.3333 0.3333])
```

```
        set(h(3),'position',[0.6666 0.6666 0.3333 0.3333])
        set(h(4),'position',[0      0      0.3333 0.3333])
        set(h(5),'position',[0.3333 0      0.3333 0.3333])
        set(h(6),'position',[0.6666 0      0.3333 0.3333])
        writeVideo(v,getframe(gcf));
        disp(i)
    end
    close(v)
end
end
%%
function dial_frame(kk, dial_number,Rotation_matrix)
plot([0 0.98*sin(Rotation_matrix(kk,dial_number))], [0 0.98*cos(Rotation_matrix(kk,dial_
number))], '--','color',[.5 .5 .5],'Linewidth',1);
hold on
xx=linspace(0,2*pi,75);
plot(0.98*cos(xx),0.98*sin(xx),'k-');
plot(0,0,'.k', 'MarkerSize',70)
ax = gca;ax.XLim = [-1.02 1.02];
set(gca,'XTick',[]);
set(gca,'YTick',[]);
axis square
axis off
end
```

Appendix B: Videos

The seven videos are available online: http://doi.org/10.4121/uuid:63affb79-d408-4f5b-9b79-8238dd42fa76

Appendix C: MATLAB script for calculating effort levels

```
[~,t,BP,AM,BPt]=signal_for_dial(3600,1); % create one hour of signals
BPtr=reshape(BPt',size(BPt,1)*size(BPt,2),1); % place all spacebar press times in 1 vector
Cr=reshape(repmat(transpose(1:6),1,size(BPt,2))',size(BPt,1)*size(BPt,2),1); % create vector of
corresponding dials
temp=find(isnan(BPtr));
BPtr(temp)=[];
Cr(temp)=[];
[BPtrs,b]=sort(BPtr); % vector of all spacebar press times in chronological order
Crs=Cr(b);  % corresponding dials
DO=perms(1:6); % 720 possible dial orders
DTT=NaN(size(DO,1),length(Crs));
for j=1:size(DO,1) % loop over 720 permutations of the 6 dials
    for i=2:length(Crs) % loop over all button press times
        PR = find(DO(j,:)==Crs(i-1)); % dial number of previous dial
        CR = find(DO(j,:)==Crs(i));   % dial number of current dial
        clear DT
        if sum(ismember([CR PR],[1 2]))==2
            DT=1;
        elseif sum(ismember([CR PR],[1 3]))==2
            DT=2;
        elseif sum(ismember([CR PR],[1 4]))==2
            DT=1;
        elseif sum(ismember([CR PR],[1 5]))==2
            DT=sqrt(2);
        elseif sum(ismember([CR PR],[1 6]))==2
            DT=sqrt(5);

        elseif sum(ismember([CR PR],[2 3]))==2
            DT=1;
        elseif sum(ismember([CR PR],[2 4]))==2
            DT=sqrt(2);
        elseif sum(ismember([CR PR],[2 5]))==2
            DT=1;
        elseif sum(ismember([CR PR],[2 6]))==2
            DT=sqrt(2);
```

```
        elseif sum(ismember([CR PR],[3 4]))==2
            DT=sqrt(5);
        elseif sum(ismember([CR PR],[3 5]))==2
            DT=sqrt(2);
        elseif sum(ismember([CR PR],[3 6]))==2
            DT=1;

        elseif sum(ismember([CR PR],[4 5]))==2
            DT=1;
        elseif sum(ismember([CR PR],[4 6]))==2
            DT=2;

        elseif sum(ismember([CR PR],[5 6]))==2
            DT=1;
        end
        DTT(j,i)=DT; % store transition distance for permutation j and transition i
    end
end
Effort=nansum(DTT,2); % sum of all transition distances for that particular permutation
levels=7;
sorth = sort(Effort);
h = zeros(levels,1);
k = 0:720/(levels-1):720;
k(1) = 1;
DO_eff = zeros(levels,6);
eff=NaN(levels,1);EL=eff;
for i = 1:levels
    h(i) = sorth(k(i));
    eff(i) = find(Effort == h(i));
    EL(i)=Effort(eff(i)); % Effort score for effort level i
    DO_eff(i,:) = DO(eff(i),:); % Dial configuration for effort level i
end
disp([DO_eff round(EL)])
```

Appendix D: Transition paths

We examined all 104,871 registered transition paths between two consecutively sampled dial AOIs across all 30 possible transition paths. Results showed that 63.8% of transition paths involved a transition from one position to another immediately above, below, left, or right of the current position, which is greater than an equal chance distribution of 46.7% for such transition paths (out of 14 possible transition paths). A further 21.5% concerned a diagonal transition to an adjacent dial, which is less than an equal chance distribution of 26.7% (e.g., from the top left to the bottom middle; 8 possible transition paths). 8.9% of transitions ran horizontally from a left/rightmost to the other left/rightmost dial, which is less than an equal chance distribution of 13.3% (4 possible transition paths), and finally 5.8% of transitions were between non-adjacent diagonals which is less than an equal chance distribution of 13.3% (e.g., from the top right to the bottom right; 4 possible transition paths). An overview of the results for all 30 transition paths is provided in Table D-I.

Appendix E: Performance scores per dial (*N* = 86)

Dial	**Performance score (%)** ***M* (*SD*)**
Top Left	46.41 (8.70)
Top Middle	42.70 (8.72)
Top Right	40.61 (11.73)
Bottom Left	52.52 (9.75)
Bottom Middle	43.56 (9.17)
Bottom Right	52.91 (9.18)
0.03 Hz	30.96 (11.13)
0.05 Hz	48.22 (11.36)
0.12 Hz	46.55 (8.68)
0.20 Hz	50.95 (9.04)
0.32 Hz	52.21 (8.45)
0.48 Hz	49.80 (8.03)

CHAPTER 3

On Senders's Models of Visual Sampling Behavior

Eisma, Y. B., Hancock, P. A., & De Winter, J. C. F. (in press). On Senders's models of visual sampling behavior. *Human Factors*.

ABSTRACT

Objective. We review the sampling models described in John Senders's doctoral thesis on 'Visual scanning processes' via a ready and accessible exposition.

Background. John Senders left a significant imprint on Human Factors/Ergonomics (HF/E). Here, we focus on one preeminent aspect of his career, namely visual attention.

Methods. We present, clarify, and expand the models in his thesis through computer simulation and associated visual illustrations.

Results. One of the key findings of Senders's work on visual sampling concerns the linear relationship between signal bandwidth and visual sampling rate. The models that are used to describe this relationship are the periodic sampling model, the random constrained sampling model, and the conditional sampling model. A recent replication study that used results from modern eye-tracking equipment showed that Senders's original findings are manifestly replicable.

Conclusions. Senders's insights and findings withstand the test of time and his models continue to be both relevant and useful to the present and promise continued impact in the future.

Application. The present paper is directed to stimulate a broad spectrum of researchers and practitioners in HF/E and beyond to use these important and insightful models.

Keywords: Visual Attention, Computer Simulation, Sampling, Replication, Bandwidth

Précis: This paper presents and explains the visual attention models of the late John Senders in an accessible form. The results of a replication study are also presented. Finally, we discuss why Senders's propositions continue to exert on-going impact.

INTRODUCTION

John W. Senders (1920–2019) left a significant imprint on a number of different areas of Human Factors/Ergonomics (HF/E), including modeling of human error, mental workload, and manual control. Herein, we focus on what is arguably the most impactful element of his scientific career: the topic of visual attention. We review and expand upon the ideas expressed in Senders's (1983) doctoral thesis entitled '*Visual Scanning Processes*.' Senders obtained his formal doctoral qualification later in life with an advisor who had once been his advisee (see Hancock et al., in press). This thesis reported a culmination of a number of his previous publications including *'The human operator as a monitor and controller of multi-degree of freedom systems'* (Senders, 1964), *'A re-analysis of the pilot eye-movement data'* (Senders, 1966), and *'The attentional demand of automobile driving'* (Senders et al., 1967).

In his thesis, Senders described experiments in which he asked participants to view a bank of dials with randomly moving pointers. This experimental configuration was inspired by the paper *'Eye movements of aircraft pilots during instrument-landing approaches'* by Fitts, Jones, and Milton (1950)**.** In the latter work, Fitts and his colleagues had concluded that the frequency of eye glances to a particular instrument indicated the relative importance of that instrument. One particular limitation of the Fitts et al. study, however, is that it is entirely descriptive, without using a quantitative model. Senders (2016) noted that "*psychologists generally shy away from seeing integral signs and partial derivatives in a paper; they just don't read it*" (50:48). In his thesis, he further explained that "*The Pilot Eye Movement Studies were being carried out but in a quite non-analytic way. That lack of analyticity was displeasing."* (Senders, 1983, p. 100). Senders, while reflecting on three distinct disciplines in human-machine systems research: human factors, engineering psychology, and the engineering approach, noted: *"I've been involved in all three"* (Senders, 2016, 50:43; Hancock et al., 2019). In his thesis, Senders sought to link the different disciplines by introducing a mathematical approach and notation concerning the psychological problem of attention distribution. Through his experiments, he demonstrated a common principle of attentional demand; namely that visual sampling rate towards any one source is linearly related to the bandwidth expressed by that specific source.

In the 1960s and 1970s, Senders's work sparked a number of follow-up modelling efforts that attempted to extend and refine his work (Carbonell, 1966; Carbonell, Ward, & Senders, 1968; Sheridan, 1970; Kvålseth, 1978). As of today, Senders's observation, and the principle of attentional demand in particular, is more broadly recognized as one of the landmark results in HF/E. Sheridan (2017), in his review of the most impactful HF/E models, identifies Senders's principle of visual sampling as being amongst them. Similarly, in a more recent review on human performance modeling, Li, Huang, and Feng

(2020) categorized Senders's findings alongside other ground-breaking models such as the Hick-Hyman Law of reaction time and Fitts' Law concerning the speed and accuracy of human movement. Wickens (2008) subsequently developed a now well-known model of visual information sampling called the Salience, Effort, Expectancy, Value (SEEV) model. For the development of the 'expectancy' component of this model, Wickens relied extensively on Senders's pioneering work. More specifically, Wickens defined the element of expectancy directly in terms of bandwidth or 'event rate.'

With respect to the above observations, it is reasonable to conclude that Senders's work is highly regarded in HF/E as well as in scientific areas beyond. However, as with many classics, it may be that it is more often cited than actually read. With certain exceptions (e.g., Moray, 1986), Senders's work on visual sampling is frequently cited in more of a 'generic' manner without referring to either its assumptions or its mathematical intricacies. Senders's work presents various equations, but hardly any visual illustrations or graphic examples that would make his work more immediately accessible to a broader audience. Another concern is that Senders performed his experiments using only a small number of participants (typically about 5). In the wake of the replication crisis in psychology (Loken & Gelman, 2017), this small number has now become some cause for concern. Accordingly, there now emerges a contemporary need for an accessible yet critical review of Senders's models, his assumptions, and his results. Herein, we look to explain and clarify Senders (1983) exposition of his models through relevant computer simulations and elaborative visual illustrations. We strive to facilitate an understanding of these models so that a wider spectrum of researchers in HF/E and beyond will be able to use them in their own research endeavors.

Task and Pointer Signals

As noted, Senders asked his participants to watch a bank of dials, each dial containing a pointer that moved in a pre-defined manner. Figure 1 illustrates this experimental setup, consisting of a number of separate booths. Participants were to press a button when any of the pointers exceeded a pre-determined threshold value on either side of the dial. Although Senders used a small number of participants, he did record extensive data from them across numerous sessions, thus: "*three minutes of camera time were obtained at the beginning and at the end of each one-hour session. There was one session on each of 30 successive days.*" (Senders, 1983, p. 42).

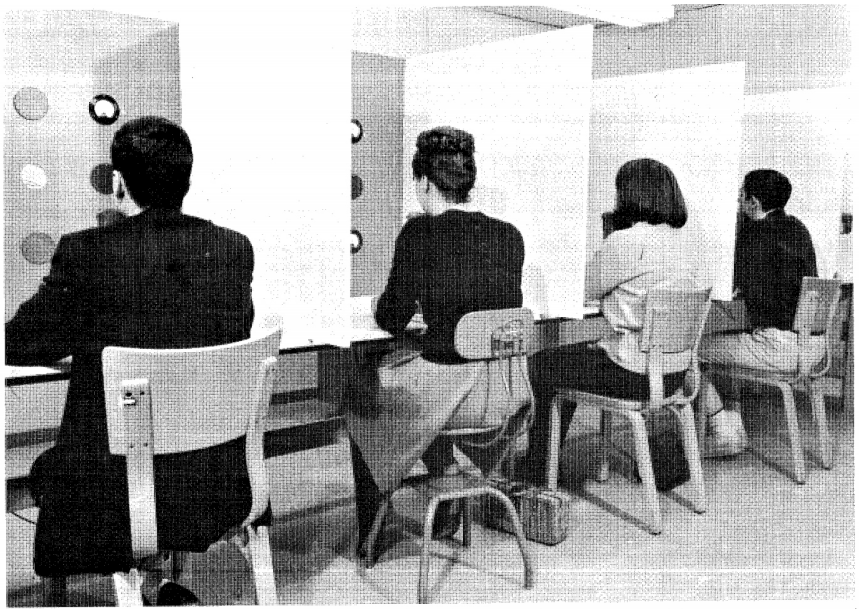


Figure 1. Experimental setup used by Senders (Senders, Elkind, Grignetti, & Smallwood, 1966).

The signals that drove the pointers, which were different for each dial, were defined using a technique that makes pointer deviations subjectively unpredictable, while its actual overall movement characteristics are known. This technique, which has also been used in manual control research, involves the summation of sine waves of different frequencies with random phase shifts (e.g., McRuer & Jex, 1967). The summation technique used to obtain the pointer signal of dial *i* as a function of time, $y_i(t)$, is described in Equation 1, in which *k* is the current sinusoid number, *m* is the number of sinusoids, Ω_k is the frequency of the sinusoid in Hz, *t* is time, and θ_k is a random phase shift for sinusoid *k*.

$$y_i(t) = \sum_{k=1}^{m} \sin(2\pi\Omega_k t + \theta_k) \quad (1)$$

Figure 2 shows a representation of two signals (*m* = 2), having frequencies of 0.03 Hz and 0.48 Hz, respectively. It can be seen that the sum of the two signals has a predictable waveform. However, by sequentially increasing the number of signal components *m*, the signal becomes progressively more unpredictable to the human observer. The signal in Figure 3 represents the summation of 1,000 sine waves, with frequencies between 0.001 and 0.48. The summed signal of dial *i* is said to have a 'bandwidth' or cutoff frequency of 0.48 Hz, W_i = 0.48 Hz, meaning that the signal is composed of a 0.48-Hz-wide 'band' of frequency components.

In this example, the frequencies were spaced linearly between 0.001 and 0.48 but another option is to use logarithmic spacing (Damveld, Beerens, Van Paassen, & Mulder, 2010; Eisma, Cabrall, & De Winter, 2018). We opted for 1,000 sinusoids to illustrate that the concept of multi-sine creation works even when the number of sinusoids is extremely high, and to ensure that all frequencies are appropriately represented. In comparison, Senders (1983) reported that "*each meter was driven by a signal, as used by Elkind (1956), composed of more than 40 sinusoids*" (p. 57). Elkind (1956), in turn, reported summing between 40 and 144 sinusoids, and claimed: "*Although 40 (the smallest number of components used frequently in these experiments) is not a very large number, it is large enough so that no periodicities in the signals are obvious*" (p. 12).

In his experiments, Senders presented participants with either four or six identical dials, each having a different signal bandwidth. A low-bandwidth signal appears as a slowly moving pointer and should hypothetically require little relative attention for detecting critical events, i.e., detecting whether the pointer angle exceeds the threshold value. A high-bandwidth pointer, it was hypothesized, demands more attention.

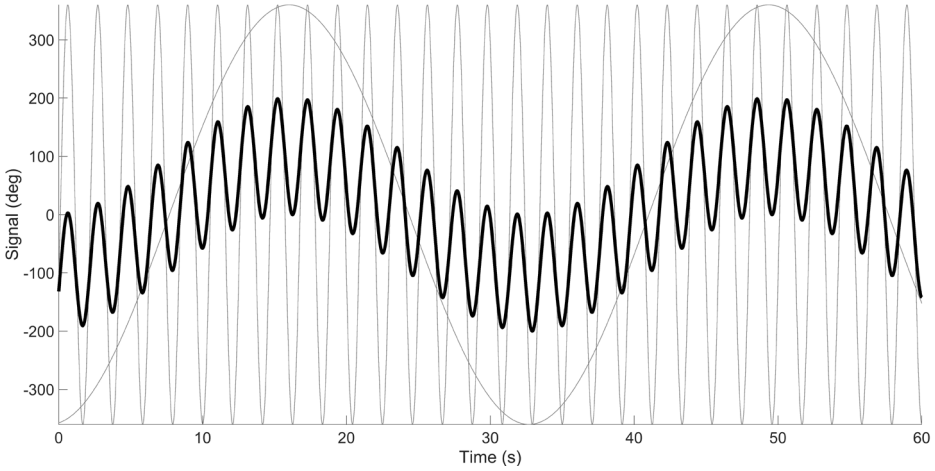


Figure. 2. Two sinusoids (in grey) and the summed signal (in black). The summed signal has been divided by a constant so that the standard deviation equals 100 deg.

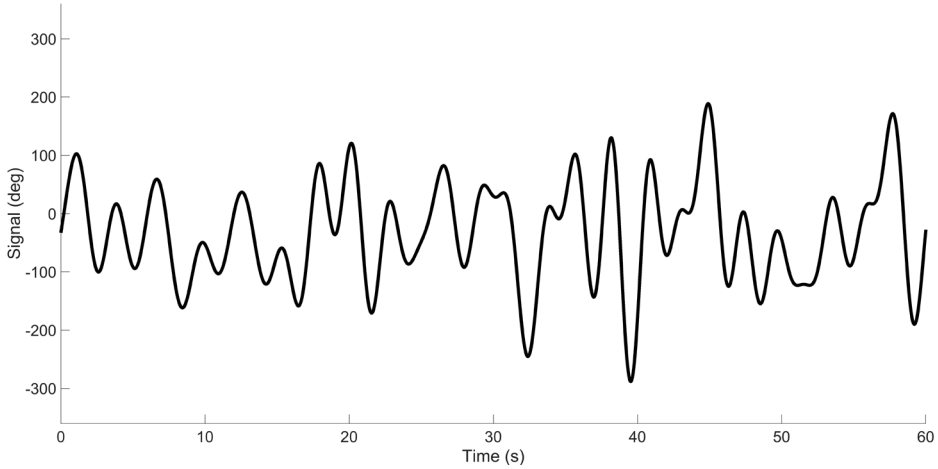


Figure. 3. Signal having a bandwidth of 0.48 Hz. The signal is the sum of 1,000 sinusoids. The summed signal has been divided by a constant so that the standard deviation equals 100 deg.

Nyquist-Shannon Sampling Theorem

Senders was concerned with predicting how often an 'ideal human observer' ought to visually sample each of the dials. He explains his Eureka moment that led him to a solution: "*upon re-reading Shannon (3) … I experienced a sudden awareness of the significance of the sampling notion for the understanding of the visual scanning behaviour of human beings*." (Senders, 1983, p. 100). More specifically, Senders conceived of adapting the Nyquist-Shannon sampling theorem for predicting human visual sampling behavior. The Nyquist-Shannon Sampling Theorem states that if one were to seek to sample a signal without losing information, one must sample (i.e., take an observation) at a frequency that is at least twice the bandwidth (i.e., the highest frequency in the signal). This theorem is illustrated in Figure 4, which shows, in black, a signal consisting of two sinusoids having frequencies of 0.03 Hz and 0.48 Hz, respectively. If one samples at 0.3 Hz, that is, if taking a reading only 0.3 times per second, it is effectively impossible to reconstruct, and thus accurately respond to the signal. The situation does not improve when sampling at 0.5 Hz or even 0.75 Hz, but when sampling at 1 Hz, which is more than twice the highest frequency (0.48 Hz) of the signal, then perfect signal reconstruction becomes possible. In these simulations, the signal was reconstructed using the Whittaker-Shannon interpolation formula (Matthé, 2017).

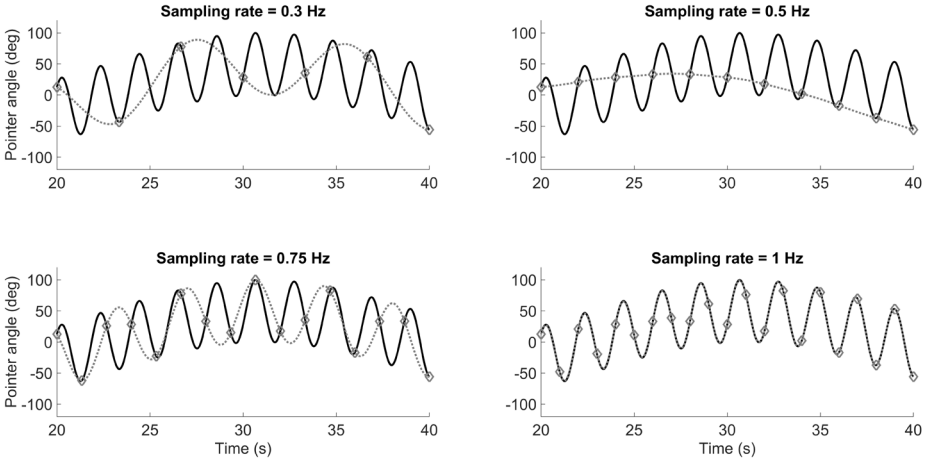


Figure 4. Illustration of the Nyquist-Shannon sampling theorem. If sampling at a frequency that is more than twice the bandwidth, signal reconstruction is possible. The red dotted line represents the signal that is reconstructed from the samples. The samples are represented by the markers.

Periodic Sampling Model

Thus, Senders hypothesized that human operators behave as ideal observers who attempt to reconstruct the observed signals. To do this, the human would have to periodically sample the signal, just as shown in Figure 4. We should caution that Senders did not believe that humans actually act as periodic samplers: *"The periodic sampler was originally constructed as a simple and unrealistic model of human behaviour"* (Senders, 1983; p. 37; emphasis ours). That is, it is rather unlikely that humans act to formulate perfect knowledge of the signal characteristics and then sample at a fixed frequency that is entirely independent of the momentary pointer angle and its velocity. However, Senders was sanguine about his use of the Nyquist-Shannon Theorem as a foundation for his models. He then produced evidence that human operators do behave in strong accordance with such a sampling criterion. Figure 5 (redrawn from Senders's thesis, p. 51) shows that participants actually sampled at a frequency (2.44*W*, *W* being the bandwidth of the pointer signal), just above the Nyquist rate (2W).

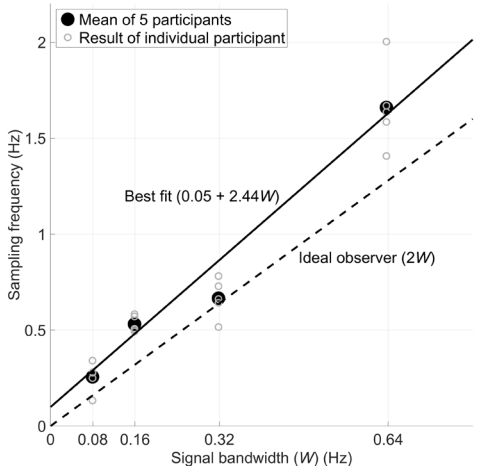


Figure 5. Senders's results show a convincingly strong correlation (*r* = .98 at individual and aggregate levels) between bandwidth and visual sampling. Five participants watched four dials and pressed a button when any of the four signals exceeded a threshold angle. Eye movements were recorded using a motion-picture camera.

In his Periodic Sampling Model (PSM), Senders defined the amount of attention that each dial required. More specifically, the attentional demand (T_i of a particular dial was defined as two times the product of the signal bandwidth (W_i) and the sampling duration (D_i) for that dial:

$$T_i = 2W_i \times D_i \tag{2}$$

For example, if the signal bandwidth was 0.48 Hz, and the sampling duration is held constant at 0.30 s, which would be a typical fixation duration (e.g., Rayner, 1998), then T_i is 0.288. This would mean that, if a sampling task lasts, for example, 100s, dial *i* would be expected to absorb 28.8s of attention.

The attentional demand for all dials combined is:

$$T_{sum} = \sum_{i=1}^{m} T_i \tag{3}$$

where *m* specifies the number of dials:

Using Equation 3, it is eminently possible to then guide the design of the human-machine interfaces. For example, if there are currently *m* dials (or displays/tasks) and there arises the need to add another (*m*+1)-th, this becomes possible only if $T_{sum} + T_{m+1} \leq 1$. Otherwise, adding that demand overwhelms the human observer.

The attentional demand as computed in Equations 2 and 3 represent ideal values. In reality, humans prove to be rather less efficient, and the combination of signal bandwidths may inhibit periodic sampling. As explained by Senders, it is unlikely that periodic sampling is feasible when there is more than one dial: "*Only in extremely rare circumstances would it be possible for strictly periodic sampling to take place on a multitude of instruments in an operational task. The periods would have to be commensurable and of such size as to permit a repeated sequence to occur*" (p. 28).

Random Constrained Sampling Model

In response to the above observation, Senders proposed an alternative model, which he called the Random Constrained Sampling Model (RCM). This model proves similar to the PSM since it assumes that the human operator samples the dials according to their bandwidths. However, the RCM assumes that the human is otherwise ignorant and samples the instruments entirely randomly instead of periodically. In the RCM, the probability that instrument *i* is sampled (p_i)equals the bandwidth of the signal relative to the total bandwidth of all signals. For example, if the human operator is looking at four dials, having bandwidths of 0.08, 0.16, 0.32, and 0.64 Hz, then the sampling probabilities, p_i, of those four dials would be 6.7, 13.3, 26.7, and 53.3%, respectively (Equation 4).

$$p_i = \frac{W_i}{\sum_{i=1}^{m} W_i} \tag{4}$$

Predicting the sampling interval for a particular dial *i*, under the assumption that each dial is sampled independently, can be accomplished using the geometric distribution shown in Equation 5. Here, p_i is the probability that a sample of the human operator falls on a dial *i* (as defined in Equation 4), and $P_r(X_i = k)$ denotes the distribution of the probability that the *k*-th sample (*k* = 1, 2, 3, …) of the human operator falls on dial *i* for the first time.

$$P_r(X_i = k) = (1 - p_i)^{k-1} p_i \tag{5}$$

For example, suppose that p_i = 13.3%, then the probability is 13.3% that the human operator first samples dial *i* on the first occasion (*k* = 1), the probability is 11.6% that the human operator first samples dial *i* on the second occasion (*k* = 2), the probability is 10.0% that the human operator first samples dial *i* on the third occasion (*k* = 3), etc. This distribution converges to 0 for large *k*, because it becomes increasingly likely that dial *i* has already been sampled.

If we assume a fixed sampling duration D for all dials (an assumption that can be justified according to Senders, because the required precision of reading is the same for all dials, Senders, 1983, p. 29), then the expected value of the time interval between two successive fixations on any particular dial becomes (see e.g., Dekking et al., 2005, p. 153):

$$\mu_{interval,i} = 1Dp_i(1-p_i)^0 + 2Dp_i(1-p_i)^1 + 3Dp_i(1-p_i)^2 + \cdots = \frac{D}{p_i} \quad (6)$$

For example, if a dial is sampled with a probability of 13.3% (p_i = 13.3%) and the sampling duration D is 0.3 s, then the expected sampling interval for dial *i* is 2.25 s. An illustration of these sampling efforts is provided in Figure 6; shown here for a four-dial instrument panel. In this illustration, Dial 2 is sampled on average every 2.25 s.

However, Senders (1983) also made predictions about the sampling duration of each instrument. Up until this point, we have assumed a fixed sampling duration, *D*. However, as shown in Figure 6, due to occasional repeated sampling of the same dial, the observed *D* proves to be higher than 0.3 s. The correction factor used by Senders is specified in Equation 7.

$$\overline{D_o} = D\frac{1}{1-p_i} \quad (7)$$

The result of this correction results for Dial 2, for example, in the average observed fixation duration,D_o, to be 0.346 s rather than of 0.3 s.

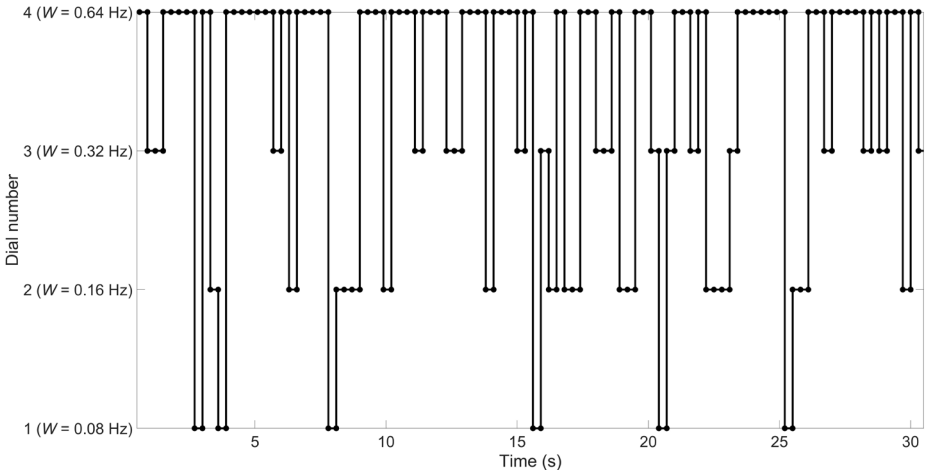


Figure 6. Example of 30 s of random sampling for a fixation duration of 0.3 s sampling probabilities of 6.7, 13.3, 26.7, and 53.3% for Dials 1, 2, 3, and 4, respectively.

Conditional Sampling Model

The RCM assumes that the human samples the dials based on bandwidth only. This would imply that the human, after practice, has formed some type of mental model of the bandwidths of the dials. Senders (1983, pp. 85–86) stated that: "*The decision for the practiced observer is hardly in the nature of a voluntary one. … Rather it is as if the eyes' mind, earlier hypothesized, directed the eyes in such a way as to bring to attention what the mind's eye wanted to see*."

In practice, however, human operators may sample a dial not only based on bandwidth but also based on the absolute position of the pointer and/or its velocity. Senders (1983, p. 32), in typical fashion, illustrated this through the use of a metaphor of a baby crawling in the vicinity of a swimming pool: *"Imagine yourself to be trying to read this monograph whilst seated on the lawn near a swimming pool. An infant is crawling on the grass generating a 'random crawl'. You wish to intervene when the infant is likely to fall into the pool and you wish to get as much reading done as possible, as well. A sensible strategy would be to calculate a next time to look at the infant based on what you had observed on the last look. If the infant had been close to the pool's edge, you would look much sooner than if it had been far away. Other things being equal, you would look sooner if it had been approaching the edge of the pool than if it had been receding from it. Lastly, in general and other things being equal, you would look sooner if it were an active infant than if it were a lethargic one. Thus the determinants of your observing behaviour would be: the amount by which the value observed fell short of the limit; the derivative of the observed variable; and the mean absolute velocity of the variable (which will be a direct function of the bandwidth of the signal formed by the positions of the infant in time)."*

As a consequence, in his thesis, Senders thus proposed a Conditional Sampling Model (CSM). The essence of this is that the human operator samples the dials based on uncertainty. A key variable in this model is the autocorrelation of the signal displayed by the pointer, that is, the correlation of the signal with a time-delayed copy of itself. The autocorrelation ρ equals 1 if the time shift τ is 0 s. The autocorrelation theoretically drops with increasing τ. The idea of this time-dependent uncertainty is illustrated in Figure 7. Suppose that, at a given moment, *t*, the human observes dial *i* having a pointer angle of 70 deg $(y_i(t) = Y_i = 70\ deg)$. If the human resamples dial *i* shortly afterwards, for example, at *t* + τ = 0.1 s, then the pointer is most probably still close to that 70-deg. value. The expected value of the pointer angle with respect to time, given an initial reading Y_i, is provided by Equation 8, where $\rho_i(\tau)$ is the autocorrelation function of the signal (Senders, Elkind, Grignetti, & Smallwood, 1966, p. 23). It is noted that Equation 8 assumes that the operator reads only the current pointer angle, Y_i, not the pointer velocity. Better predictions will be possible, and the operator will therefore need to sample less often, if also reading pointer velocity (Fogel, 1955).

$$\mu_{\hat{y}_i(t+\tau)|y_i(t)=Y_i} = \rho_i(\tau)Y_i \tag{8}$$

For a bandlimited white noise signal with linearly spaced frequencies (see Figure 3 above), the autocorrelation function, $\rho_i(\tau)$, is known to be (Senders, 1983, p. 33; Knudtzon, 1949, p. 6):

$$\rho_i(\tau) = \frac{\sin(2\pi W_i \tau)}{2\pi W_i \tau} \tag{9}$$

So, suppose the signal bandwidth W_i is 0.48 Hz, then the expected value of the pointer angle for τ = 0.1 s will be 68.9 deg, or very close to the original 70 deg. As τ increases, the initial reading of 70 deg becomes less and less influential. For example, if τ = 1.0 s, then the expected value is 2.9 deg, that being much closer to the overall expected value of 0 deg (see the solid line Figure 7, left).

The standard deviation of the predicted the future signal, $\hat{y}_i(t+\tau)|y_i(t) = Y_i$, can be computed using Equation 10 (Senders, 1983, p. 34). This standard deviation is a measure of the uncertainty of the prediction, where it is noted that $1 - \rho_i^2(\tau)$ is the fraction of the variance in the predicted future signal that is uncorrelated with $y_i(t)$.

$$\sigma_{\hat{y}_i(t+\tau)|y_i(t)=Y_i} = \sigma_{y_i}\sqrt{\left(1-\rho_i^2(\tau)\right)} \quad (10)$$

At the moment of sampling the dial, the human is entirely certain about the status of the signal. That is, $\sigma_{\hat{y}_i(t+\tau)|y_i(t)=Y_i}$ = 0 deg. The longer the operator does not sample the dial, the higher the uncertainty. If τ is 0.1 s and the standard deviation of the entire signal σ_{y_i} is 100 deg, then $\sigma_{\hat{y}_i(t+\tau)|y_i(t)=Y_i}$ is 17.3 deg, which is effectively small. The standard deviation rises to its nominal value of 100 deg as τ increases. This is illustrated in Figure 7, left, in which the standard deviation of the predicted pointer angle is the distance between the solid mean line and the two dotted lines. The rise of uncertainty about the value of the signal is analogous to the metaphorical crawling infant. If the parent has not seen the infant for a while, then logically, the parent should be ever-more concerned about whether the infant has fallen into the pool.

Let us now assume that the operator is tasked to press a button whenever the pointer exceeds a critical value L_i. We can, from the foregoing observations, calculate the probability that this pointer angle exceeds the critical value (one-sided), using the normal cumulative distribution function (Φ, as shown in Equation 11 (Senders et al., 1966, p. 37; Senders, 1983, p. 34). For example, assume, L_i = 100 deg, Y_i = 70 deg, and σ_{y_i} = 100 deg. Then, at τ = 0.1 s, the probability of the pointer exceeding the 100-deg critical value is 3.6%, (see Figure 7, right). In other words, one-tenth of a second after the initial reading, the probability is only a relatively small one that the pointer is actually exceeding the threshold. Given this understanding, the operator may be disinclined to sample the dial again.

$$P_{\hat{y}_i(t+\tau)>L_i|y_i(t)=Y_i} = 1-\Phi\left(\frac{L_i-\mu_{\hat{y}_i(t+\tau)|y_i(t)=Y_i}}{\sigma_{\hat{y}_i(t+\tau)|y_i(t)=Y_i}}\right) = 1-\Phi\left(\frac{L_i-\rho_i(\tau)Y_i}{\sigma_{y_i}\sqrt{\left(1-\rho_i^2(\tau)\right)}}\right) \quad (11)$$

Senders (1983) proposes various types and forms of CSMs in his thesis. For example, in CSM-1, the operator is assumed to sample the dial when the probability of exceeding the critical value is maximal (0.47 s in Figure 7). In CSM-2, the operator samples the dial when the probability of exceeding the critical value is greater than any particular specified probability threshold. The essence of CSM models is that the sampling frequency does not depend on bandwidth alone (W_i, i.e., whether the infant is lethargic or not) but also on the last pointer reading (y_i; thus, where the infant was in relation to the swimming pool when last seen).

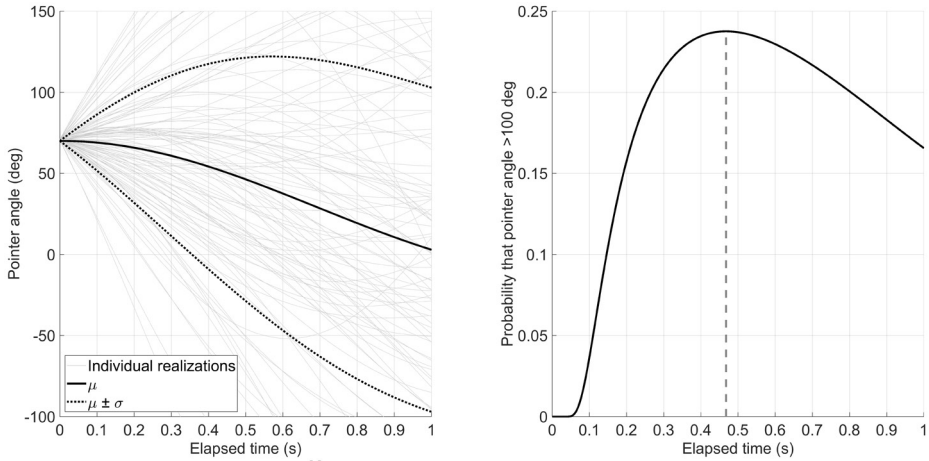


Figure 7. Left: Expected value, $\mu_{\hat{y}_i(t+\tau)|y_i(t)=Y_i}$ and corresponding standard deviation of the predicted future signal, $\sigma_{\hat{y}_i(t+\tau)|y_i(t)=Y_i}$ for elapsed time (τ) since taking a reading Y_i of 70 deg. Also shown are a random 100 realizations of the pointer angle with $Y_i \approx 70$ deg. Right: The corresponding probability that the pointer angle exceeds 100 deg; the probability has a maximum at $t + \tau = 0.47$ s. The pointer signal has a 0.48 Hz bandwidth and overall standard deviation of 100 deg (see Figure 3 for an example of the signal).

Modern Replication of Senders's Experimental Findings

As we noted previously, Senders (1983) used only small numbers of participants (see Figure 5). In line with the recent popularity of replication research (Zwaan, Etz, Lucas, & Donnellan, 2018), Eisma et al. (2018) performed just such a replication and expansion of Senders's 6-dial study, using Senders's own specified bandwidths of 0.03, 0.05, 0.12, 0.20, 0.32, and 0.48 Hz. The replication study was however conducted using 86 total participants. The results revealed a remarkably strong and gratifying congruent outcome with Senders's original results. More specifically:

- Both Senders (1983) and Eisma et al. (2018) found that participants' mean sampling rate was proportional to bandwidth. More specifically, Eisma et al. (2018) found the following best fit: *Sampling rate* = $0.64W + 0.20$ (r = .98), whereas Senders had earlier found that. *Sampling rate* = $0.61W + 0.18$ (r = .99) Note that the slope of approximately $0.61W$ is considerably shallower than the slope of $2.44W$, as observed for the 4-dial task (Figure 5). This, we believe, is because the 6-dial task was attentionally more demanding, the result of which was that participants were unable to distribute their attention optimally.
- Eisma et al. (2018) found an increase of glance duration as a function of dial bandwidth. Senders had originally predicted this effect (see Equation 7), but the

empirical data in Senders's thesis were indeterminate here, again, perhaps due to the reliance on a small number of participants and the low eye-movement data acquisition rate (12 fps) of his equipment's measurement capacities at that time. Thus, we have not simply confirmed some of Senders's original findings but have established the veracity of one of his predictions that, due to inherent constraints, he was not able to fully evaluate in this own experimentation. More specifically, Eisma et al. found the following best fit: $Mean\ glance\ duration$ = 0.43W + 0.38 (r = .99), whereas Senders had earlier reported a weaker fit via the specification that *Mean glance duration* = 0.30W + 0.47 (r = .81).

- Senders (1983) was unable to fully test the CSM, but manually annotated participants' eye-movements, based on camera images recorded. He had no way of relating these camera recordings directly to the current pointer angle. The replication study of Eisma et al. (2018) used modern eye-tracking equipment together with synchronous data recordings in order to be able to accomplish this. The latter authors found, in agreement with the crawling infant analogy, that (1) bandwidth, (2) pointer angle, i.e., how close to the threshold it was, and (3) pointer velocity (higher velocities attracting more attention) each strongly influenced the probability that a participant then glanced at a specific dial.

  In addition to these encouraging findings, Eisma et al. (2018) noted several points where Senders (1983) may have potentially erred or provided only an incomplete explanation.

- Senders did not specify how the dials were arranged on the instrument panel. For example, if the high-bandwidth dials were placed in the middle, then the associated visual sampling effort might be relatively low. In contrast, if the high-bandwidth dials were positioned toward the edges, then visual sampling effort could have been relatively higher since the human operator would then need to scan across greater distances. Eisma et al. (2018) found that this 'effort configuration' does matter, with less ideal sampling (i.e., lower than the Nyquist rate of $2W$) when the needed effort level was higher.

- Senders claimed that people need extensive practice: "*So I trained my subjects for more than 30 hours and took data along the way in order to find out how long it took for them to stabilize in their scanning behaviour. Indeed, it took about ten hours for scanning to stabilize and more nearly twenty-five for detection to arrive at a reasonable high level*." (p. 101). Eisma et al. (2018) found, however, that after only 20 s on the task, a clear distinction appeared between the sampling rates for the different dials. This finding suggests that conditional sampling, that is, sampling based primarily or even exclusively on bottom-up sensory cues, represents the dominant psychological mechanism employed by operators.

- Senders (1983) required operators to detect threshold crossings but did not report on any performance data per se. De Winter, Eisma, Cabrall, Hancock, and Stanton (2019) showed a strong correlation between participants' sampling behavior and their detection performance (r = .78). In other words, people who showed superior sampling (i.e., looking at the right dial at the right time) detected more threshold crossings. Although this might appear to be even self-evident, such a clear link between the spatial orientation of attention and subsequent detection efficiency appears not to have been demonstrated before. For, elsewhere, it has been determined that looking (i.e., the fixation of the eyes) does not necessarily equate to seeing (i.e., the processing of information in that fixated area) (and see Krueger et al., 2019).

DISCUSSION

In this paper, we have reviewed several facets of John Senders's collective work '*Visual Sampling Processes'*. Our goal was to make this important work accessible to a broader audience. We have therefore provided illustrations of (1) how to create a random-appearing pointer signal, and what it means to (2) sample periodically in order to reconstruct that pointer signal, (3) sample randomly in a bandwidth-constrained manner, or (4) sample conditionally based on pointer value during the last sample of the dial. Furthermore, we reviewed a recent replication study which demonstrated that the findings of Senders readily replicate. The latter experimental study also validated several of Senders's predictions that he himself was unable to test with the equipment of his time. Overall, our treatise is intended to recognize Senders's legacy and to show how his ideas remain relevant to many modern applicational contexts.

We might ask why Senders's work was so readily replicable? We suggest two major reasons. First, Senders's empirical findings and untested predictions were based upon substantive models and calculations. As pointed out by Box (1976), "*all models are wrong*" (p. 792). By this he meant that all models, in their attempt to make accurate predictions, rest on assumptions and therefore, cannot predict the real world *exactly*. All models are necessarily reductions of the world that they seek to portray and so must be, at best, only reduced approximations. In the case of Senders's models, it is unlikely that humans would sample periodically or randomly without any consideration as to the state of the dials they are viewing. It is also unlikely that humans can flawlessly estimate the probability that the pointer angle would exceed a target threshold angle. Furthermore, according to the aforementioned Wickens's (2008) SEEV model, there is more to sampling than expectancy and salience alone. 'Value' (the cost of not sampling a particular dial) and 'effort' (the amount of eye-movement and head-movement required, as explained above in the modern replication of Senders) also each affect eye movements. However, regardless of these assertions and assumptions,

Senders's models do provide a plausible and useful basis for predicting human sampling behavior. Sheridan (2002) referred to these type of models as *"borrowed engineering models"*. That is, in and around the 1960s, HF/E researchers started to deviate from purely descriptive 'knobs and dials' research, such as the Fitts et al. (1950) studies, and started using the then available quantitative models. These models were, perhaps naturally 'borrowed' from the engineering domain since that discipline possessed the most relevant and applicable ones at that time. In Senders's case, this was from Shannon's work on information theory (see Shannon & Weaver, 1949). By promulgating these theoretical bases, more realistic predictions of sampling behavior can be made, as compared to purely descriptive approaches. As Senders put it: "*I went back to Shannon, the 1947 article, read the thing again, and decided that the Sampling Theorem would be the controlling factor. Irrespective of what people wanted to do, what they could do, the limitations would be mathematically defined.*"

A second reason for the high degree of replicability is that Senders did not rely on null-hypothesis significance testing. Probability estimates are nowhere to be found in his work. Instead of performing assessments predicated upon which condition yields significantly different results from comparative and control conditions, Senders estimated functional relationships between experimental variables, e.g., between sampling rate and bandwidth. In recent work, Smith and Little (2018) have explained why this type of approach to psychological research is expected to render such replicable results. They showed, through theoretical argument and computer simulation, that high replicability can be achieved even when only a small number of participants are subjected to a large number of trials. This approach corresponds to typical methods in psychophysics (Smith & Little, 2018).

Now that technology is becoming increasingly automated, the human operator is often only a supervisor rather than a direct controller (Hancock, 2014; Sheridan, 2002). For example, in automated car driving, the driver does not necessarily have to turn the steering wheel or press any of the foot pedals. However, presently, drivers still have to monitor the road and occasionally re-take vehicle control (Eriksson & Stanton, 2017; Zhang, De Winter, Varotto, Happee, & Martens, 2019). Since active manual control is absent in automated driving, there is an increasing research focus on indirect control, e.g., gestural control and monitoring. Here, the human has to monitor the automation that can, in its turn, monitor the human. Perhaps not surprisingly, in a few years, driver drowsiness/attention monitoring systems will be obligatory in newly sold passenger cars in the European Union (European Commission, 2019). We expect that Senders's work to become increasingly more relevant in such human-automation interaction on both research and application fronts. In his work, Senders used visual occlusion methods to determine the attentional demands of drivers (Saffarian, De Winter, & Senders, 2015; Senders, 1964). Instead of merely detecting whether drivers are visually attentive or

distracted, a computational model could be used to determine whether the driver has sampled the relevant objects in the driving environment (De Winter et al., 2019). We anticipate that Senders's models will here represent a useful starting point to such computational models. It may be possible, for example, to provide a warning if driver sampling behavior deviates significantly from expectations as determined from signal bandwidth. More specifically, we can postulate that drivers do not have to distribute their attention uniformly across their ambient working environment, but mainly have to look at regions in the visual field where activity is taking place. In other words, human drivers will be obliged to look a lot at the roadway and mirrors, and also at the dashboard, and much less at the scenery or parked cars. As derived from the SEEV model (Wickens, 2008), a separate computational module will have to determine which road regions have high value, and the higher values will have to be assigned to objects that have a higher probability of colliding with the driver's car.

Our final point on attention concerns the interplay between top-down (bandwidth-based) and bottom-up sampling (pointer angle, velocity) processes. We can affirm and confirm that both factors are relevant, but it remains unknown at present exactly how they jointly contribute to human sampling (and see Hancock, 2019). A high-bandwidth display may attract attention because operators perceive something moving rapidly in their peripheral vision. However, perhaps after an extended period of observation (minutes or even hours), the operator can form expectancies about where to look predicated upon their accumulated situational experience and not just the momentary dynamics of the display(s) before them (e.g., "the left top dial requires most of my attention"). The interplay between the top-down cues (expected value) and bottom-up cues (salient features such as a fast-moving dial) obviously requires further research, an endeavor that**,** we believe, Senders would be in wholehearted agreement with. Such research may be performed using a gaze-contingent sampling paradigm, in which peripheral vision is occluded. In conclusion, we have looked to give a brief encomium for, and support of Senders's models of visual attention. We anticipate that the work of Senders will remain relevant to HF/E research and in realms beyond for many years to come.

REFERENCES

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association, 71*, 791–799.

Carbonell, J. R. (1966). A queueing model of many-instrument visual sampling. *IEEE Transactions on Human Factors in Electronics, HFE-7*, 157–164.

Carbonell, J. R., Ward, J. L., & Senders, J. W. (1968). A queueing model of visual sampling experimental validation. *IEEE Transactions on Man-Machine Systems, 9*, 82–87.

Damveld, H. J., Beerens, G. C., Van Paassen, M. M., & Mulder, M. (2010). Design of forcing functions for the identification of human control behavior. *Journal of Guidance, Control, and Dynamics, 33*, 1064–1081.

Dekking, F. M., Kraaikamp, C., Lopuhaä, H. P., & Meester, L. E. (2005). More computations with more random variables. In *A modern introduction to probability and statistics* (pp. 151–166). Springer.

De Winter, J. C. F., Eisma, Y. B., Cabrall, C. D. D., Hancock, P. A., & Stanton, N. A. (2019). Situation awareness based on eye movements in relation to the task environment. *Cognition, Technology & Work, 21*, 99–111.

Eisma, Y. B., Cabrall, C. D. D., & De Winter, J. C. F. (2018). Visual sampling processes revisited: replicating and extending Senders (1983) using modern eye-tracking equipment. *IEEE Transactions on Human-Machine Systems, 48*, 526–540.

Elkind, J. I. (1956). *Characteristics of simple manual control systems* (Doctoral dissertation, Massachusetts Institute of Technology).

Eriksson, A., & Stanton, N. A. (2017). Takeover time in highly automated vehicles: noncritical transitions to and from manual control. *Human Factors, 59*, 689–705.

European Commission (2019). New safety features in your car. Retrieved from https://ec.europa.eu/docsroom/documents/34588

Fitts, P. M., Jones, R. E., & Milton, J. L. (1950). Eye movements of aircraft pilots during instrument-landing approaches. *Aeronautical Engineering Review, 9*, 24–29.

Fogel, L. J. (1955). A note on the sampling theorem. *IRE Transactions – Information theory, 9*, 47–48.

Hancock, P. A. (2014). Automation: how much is too much? *Ergonomics, 57*, 449–454.

Hancock, P. A. (2019). On the dynamics of conspicuity. *Human Factors, 61*, 857–865.

Hancock, P. A., Crichton-Harris, A., Sellen, A., Sheridan, T. B., & Hancock, G. M. (in press). *A distracted scientist: The life and contributions of John Senders*.

Hancock, P. A., Hancock, G. M., Senders, W., & Sellen, A. J. (2019). John Senders (1920–2019). *The American Journal of Psychology, 132*, 361–367.

Knudtzon, N. (1949). *Experimental study of statistical characteristics of filtered random noise* (Technical Report No. 115). Research Laboratory of Electronics, Massachusetts Institute of Technology. Retrieved from https://dspace.mit.edu/bitstream/handle/1721.1/4931/RLE-TR-115-04711221.pdf

Krueger, E., Sawyer, B. D., Chavaillaz, A., Sonderegger, A., Schneider, A., Groner, R., & Hancock, P. A. (2019). Microsaccades distinguish looking from seeing. *Journal of Eye Movement Research*, *12* (6), 2: 1-14.

Kvålseth, T. O. (1978). Human information processing in visual sampling. *Ergonomics, 21*, 439–454.

Li, N., Huang, J., & Feng, Y. (2020). Human performance modeling and its uncertainty factors affecting decision making: a survey. *Soft Computing, 24,* 2851–2871.

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*, 584–585.

Matthé, M. (2017). Formulating a function on Matlab for the Shannon interpolation formula [code]. Retrieved from https://dsp.stackexchange.com/questions/37480/formulating-a-function-on-matlab-for-the-shannon-interpolation-formula

McRuer, D. T., & Jex, H. R. (1967). A review of quasi-linear pilot models. *IEEE Transactions on Human Factors in Electronics, HFE-8*, 231–249.

Moray, N. (1986). Monitoring behavior and supervisory control. In K. Boff, L. Kaufman, and J. Thomas (Eds.), *Handbook of perception and human performance* (pp. 40:1–40:55). New York: Wiley.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*, 372–422.

Saffarian, M., De Winter, J. C. F., & Senders, J. W. (2015). Measuring drivers' visual information needs during braking: A simulator study using a screen-occlusion method. *Transportation Research Part F, 33*, 48–65.

Senders, J. W. (1964). The human operator as a monitor and controller of multidegree of freedom systems. *IEEE Transactions on Human Factors in Electronics, 1*, 2–5.

Senders, J. W. (1966). A re-analysis of the pilot eye-movement data. *IEEE Transactions on Human Factors in Electronics, 2*, 103–106.

Senders, J. W. (1983). *Visual sampling processes* (Doctoral Dissertation). Katholieke Hogeschool Tilburg, The Netherlands.

Senders, J. W. (2016). History of human factors [video file]. https://www.youtube.com/watch?v=QZH97Xy6NA8&t=2801s

Senders, J. W., Elkind, J. I., Grignetti, M. C., & Smallwood, R. (1966). *An investigation of the visual sampling behaviour of human observers.* (Rep. no. NASA-CR-434), Washington, DC, USA.

Senders, J. W., Kristofferson, A. B., Levison, W. H., Dietrich, C. W., & Ward, J. L. (1967). The attentional demand of automobile driving. *Highway Research Record, 195*, 15-33.

Shannon, C. E., & Weaver, W. (1949). The mathematical theory of communication. Urbana, IL: University of Illinois Press.

Sheridan, T. B. (1970). On how often the supervisor should sample. *IEEE Transactions on Systems Science and Cybernetics, 6*, 140–145.

Sheridan, T. B. (2002). *Humans and automation: Systems design and research issues*. Santa Monica/New York: Human Factors and Ergonomics Society/Wiley.

Sheridan, T. B. (2017). *Modeling human-system interaction: Philosophical and methodological considerations, with examples*. Hoboken. NJ: Wiley & Sons, 10, 9781119275275.

Smith, P. L., & Little, D. R. (2018). Small is beautiful: In defense of the small-N design. *Psychonomic Bulletin & Review, 25*, 2083–2101.

Wickens, C. D. (2008). Visual attention control, scanning, and information sampling. In C. D. Wickens & J. S. McCarley (Eds.), *Applied attention theory* (pp. 41–61). Boca Raton: CRC Press.

Zhang, B., De Winter, J., Varotto, S., Happee, R., & Martens, M. (2019). Determinants of take-over time from automated driving: A meta-analysis of 129 studies. *Transportation Research Part F: Traffic Psychology and Behaviour, 64*, 285–307.

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences, 41*, e120.

CHAPTER 4

Situation awareness based on eye movements in relation to the task environment

De Winter, J. C. F., Eisma, Y. B., Cabrall, C. D. D., Hancock, P. A., & Stanton, N. A. (2019). Situation awareness based on eye movements in relation to the task environment. *Cognition, Technology and Work,* 21, 99–111. Joint first authors

ABSTRACT

The topic of situation awareness has received continuing interest over the last decades. Freeze-probe methods, such as the Situation Awareness Global Assessment Technique (SAGAT), are commonly employed for measuring situation awareness. The aim of this paper was to review validity issues of the SAGAT and examine whether eye movements are a promising alternative for measuring situation awareness. First, we outlined six problems of freeze-probe methods, such as the fact that freeze-probe methods rely on what the operator has been able to remember and then explicitly recall. We propose an operationalization of situation awareness based on the eye movements of the person in relation to their task environment to circumvent shortfalls of memory mediation and task interruption. Next, we analyzed experimental data in which participants (N = 86) were tasked to observe a display of six dials for about 10 min, and press the space bar if a dial pointer crossed a threshold value. Every 90 s, the screen was blanked and participants had to report the state of the dials on a paper sheet. We assessed correlations of participants' task performance (% of threshold crossing detected) with visual sampling scores (% of dials glanced at during threshold crossings) and freeze-probe scores. Results showed that the visual-sampling score correlated with task performance at the threshold-crossing level (r = 0.31) and at the individual level (r = 0.78). Freeze-probe scores were low and showed weak associations with task performance. We conclude that the outlined limitations of the SAGAT impede measurement of situation awareness, which can be computed more effectively from eye movement measurements in relation to the state of the task environment. The present findings have practical value, as advances in eye-tracking cameras and ubiquitous computing lessen the need for interruptive tests such as SAGAT. Eye-based situation awareness is a predictor of performance, with the advantage that it is applicable through real-time feedback technologies.

1. INTRODUCTION

1.1 Situation awareness

During the last three decades, an extensive body of research has appeared concerning situation awareness (SA). Although SA was initially characterized as "the buzzword of the '90s'" (Pew 1994), the term is now firmly embedded into the vocabulary of human factors and ergonomics. The construct of SA has received "strong endorsement" (Wickens 2015, p. 90) and is regarded as valuable in the research community (Parasuraman et al. 2008). At the same time, SA has its critics (Dekker 2015; Flach 1995) and its validity has been debated (Carsten and Vanderhaegen 2015; Millot 2015).

Interest in SA can be attributed to the fact that systems have become increasingly complex and automated (Hancock 2014; Parasuraman et al. 2008; Stanton et al. 2017). Wickens (2008) explained the growing importance of SA by noting that: "This trend reflects, on one hand, the growing extent to which automation does more, and the human operator often does (acts) less in many complex systems but is still responsible for understanding the state of such systems in case things go wrong and human intervention is required" (p. 397).

According to Endsley, SA reflects the extent to which the operator knows what is going on in their environment and is the product of mental processes including attention, perception, memory, and expectation (Endsley 2000a). More formally, SA has been defined as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" (Endsley 1988, p. 792). Endsley's model of SA thus consists of three ascending levels (Endsley 2015a). Level 1 denotes the perceptual process within the dynamic environment, Level 2 concerns a comprehension of those perceived elements from Level 1, and Level 3 SA is the projection of the future status.

1.2 The use and validity of the situation awareness global assessment techt nique (SAGAT)

Endsley (2015b) noted that "much of the disagreement on SA models that has been presented ultimately has boiled down to a disagreement on the best way to measure SA" (p. 108). It is a supportable assertion that the most oftenused method to assess SA is the Situation Awareness Global Assessment Technique (SAGAT; Endsley 1988). A Google Scholar search (August 2018) using the query "situation awareness global assessment technique" yielded 1850 papers, which proved to be considerably more than the number of hits for any competitor technique (e.g., "situation awareness rating technique" yielded 708 papers and "situation present assessment method" yielded 367 papers). SAGAT is a freeze-probe technique that requires operators to memorize and report on pre defined aspects of their task environment via queries which interrogate

aspects of either perception (Level 1 SA), comprehension (Level 2 SA), or projection (Level 3 SA). The higher the score with respect to a normative 'ground truth', the higher the operator's SA is considered to be.

As pointed out by Durso et al. (2006), "one of the arguments advanced for the importance of SA is that SA is a sensitive harbinger of performance" (p. 721). It has been shown that individual differences in task performance can be predicted from SAGAT scores to some extent. For example, it has been found that SAGAT scores correlate with performance on a military planning task (r = 0.66, N = 20; Salmon et al. 2009), teamwork performance among medical trainees (r = 0.65, N = 10 teams; Gardner et al. 2017), and performance in a surgical task (r = 0.47, but two other correlations were not statistically significant from zero, N = 97; Bogossian et al. 2014, and r = 0.81, N = 16; Hogan et al. 2006). SAGAT also relates to how well pilots handled in-flight emergencies in a simulator (r = 0.41, N = 41; Prince et al. 2007), crash-avoidance performance in a low-fidelity driving simulator (r = 0.44, N = 190; Gugerty 1997), scores on a driving-based hazard perception test (r = 0.56, N about 38; McGowan and Banbury 2004), performance in submarine track management (β between − 0.02 and 0.41, N = 171; Loft et al. 2015), and performance in air traffic control (r= 0.52, N = 18; O'Brien and O'Hare 2007).

However, other studies are less positive regarding the validity of the SAGAT. Durso et al. (1998) found that SAGAT correlated only weakly with performance of air traffic controllers (β between − 0.01 and 0.24, N = 12), whereas Lo et al. (2016, p. 335) found "a general tendency across conditions for a negative relation between SA probes and multiple performance indicators" (N < 10). Similarly, Pierce et al. (2008) found that participants with higher SAGAT scores committed fewer procedural errors and violations in an air traffic control task, but these effects were not statistically significant (N about 20, $p \geq 0.08$). Similarly, Strybel et al. (2008) found no significant association between SAGAT scores and air traffic control performance (N= 13). Additionally, Cummings and Guerlain (2007) found that overall performance scores in a missile control task were not statistically significantly correlated with SAGAT scores (N = 42), whereas Ikuma et al. (2014) found no significant correlations between SAGAT scores and control room operator performance (N = 36).

We argue that the abovementioned small-sample correlations may not be statistically reliable, due to measurement error and possible selective reporting bias. According to the principle of aggregation (Rushton et al. 1983), predictive validity is increased if the predictor and criterion are averaged across multiple measurement instances. Looking at the largest sample study (Gugerty 1997), the relatively strong correlation of 0.44 could be due to the fact that SAGAT scores and performance scores were averaged across a large number of trials per person (84 or more).

From the above observations, the question arises as to whether some of the stronger predictive correlations are inflated due to common method variance. To illustrate, McGowan and Banbury (2004) observed that SAGAT scores were strongly predictive of hazard anticipation performance (r = 0.56). This strong correlation is to be expected, as the term 'hazard anticipation' is often equated with SA (Horswill and McKenna 2004; Underwood et al. 2013). McGowan and Banbury argued that the correlation could be even stronger than 0.56: "if all the probe queries were to measure projection then a higher correlation will be found". In other words, it is no surprise that responses to SAGAT queries (e.g., 'what will happen next' queries) show strong associations with scores on a hazard anticipation test; the criterion and predictor variable are conceptually similar, and no independent performance is predicted. Also, it can be questioned whether the SAGAT has additional predictive validity, also called 'incremental validity' (Sechrest 1963), with respect to standard psychometric tests, such as tests of working memory and spatial ability (Pew 1994). This topic has been previously investigated by Durso et al. (2006). In a study using 89 participants, they found that SAGAT was not a sufficiently strong predictor of air traffic control performance to enter a stepwise regression model after diverse cognitive and non-cognitive tests had been allowed to enter the model first. This led these authors to conclude that "typical cognitive measures already capture much of what offline measures contribute" (p. 731). Indeed, it is known that psychometric test scores show positive intercorrelations (Van der Maas et al. 2006), and it is plausible that operators who possess high working memory capacity will perform well on any task, and thus will perform well on the SAGAT also (Gugerty and Tirre 2000; Sohn and Doane 2004). In other words, a statistical association between SAGAT scores and task performance may be due to a common cause such as general intelligence (g) rather than anything that is necessarily situational.

1.3 Aim of this study

As indicated above, the SAGAT is a widely used freeze-probe technique. SAGAT scores appear to be moderately correlated with task performance, while incremental validity is contentious. At present, it is unknown why the SAGAT has imperfect validity with regard to task performance. Accordingly, the research question that this paper sets out to answer is: "What are the limitations of SAGAT?", and secondly: "Is an alternative bodybased measure of SA more predictive of task performance than a freeze-probe method?" More specifically, we propose here that SA can alternatively be operationalized via eye movements of the operator in relation to the task environment.

The idea of using eye- trackers for inferring SA is not a new one per se. In their work, "Development of a novel measure of situation awareness: The case for eye movement analysis", Moore and Gugerty (2010) found that the higher the percentage of time air traffic controllers fixated on important aircraft, the higher their task performance and SAGAT performance. Our present work aims to follow up on this type of analysis

by focusing on eye movements in a dynamic environment. We postulate that eye movements reflect the extent to which an operator exerts a grip on the current environment (cf. Merleau-Ponty 1945) as part of the perception-action cycle (Neisser 1976), thus also being a predicate of task performance. In order to establish the concept of SA by means of eye movements and task relations, we have included the results of an experiment with 86 participants who performed a visual monitoring task of a dynamic system. We examined the correlations between a freeze-probe method and eye-based SA on the one hand, and task performance, on the other.

2. PROBLEMS WITH SAGAT

When using SAGAT, the ongoing task is frozen and the simulation screen is blanked out. The operator then answers queries about the task environment. SAGAT queries need not necessarily be textual (see Endsley 2000b, for a review). An example of non-textual queries is the work of Gugerty (1997) in which participants had to pinpoint the location of cars in a top-down view of the simulated road.

Six problems arise from the SAGAT, and they can be considered common to all freeze-probe techniques: (1) memory decay/bias, (2) task resumption deviations, (3) removal from the ongoing task, (4) explicit representations, (5) intermittency, and (6) non-situated cognition.

First, there is an inherent and inevitable time delay between the moment of freezing and the subsequent completion of all the required queries. This makes such measurements susceptible to memory decay and the biases associated with it. Thus, the most immediate and familiar situational features are remembered best (and these do not necessarily reflect those with the greatest task relevancy). Gugerty (1998) found that "information was forgotten from dynamic spatial memory over the 14 s that it took participants to recall whole report trials" (p. 498).

Second, after the simulation freezes, participants have to subsequently resume the task, and so post-freeze task performance and SA almost certainly deviate from non-interrupted task performance. It has been argued by Endsley (1995) that these two problems may not be fatal to measuring SA; she empirically found that the length of the time interval and task interruption have only minor effects on SAGAT scores. McGowan and Banbury (2004), on the other hand, found a negative effect of SAGAT interruption on task performance as compared to the same task without interruption.

Third, as most researchers in general seem to agree that SA refers to "the level of awareness that an individual has of a situation" (Salmon et al. 2008, p. 297 *awareness*, the experience of awareness should ideally be reflected in the nature and character of the measurement method(s) themselves (Smith and Hancock 1995). How people

respond to paper and pencil SAGAT queries, however, is only an indirect reflection of their phenomenological awareness, because they are removed from the situation by blanking the screen and interrupting the ongoing flow of behaviour. The task of completing SAGAT queries is temporally (i.e., the operator completes queries every few minutes while the simulator is frozen) and functionally (i.e., the operator completes queries by means of a pencil, keyboard, or touchscreen) separate from the actual task.

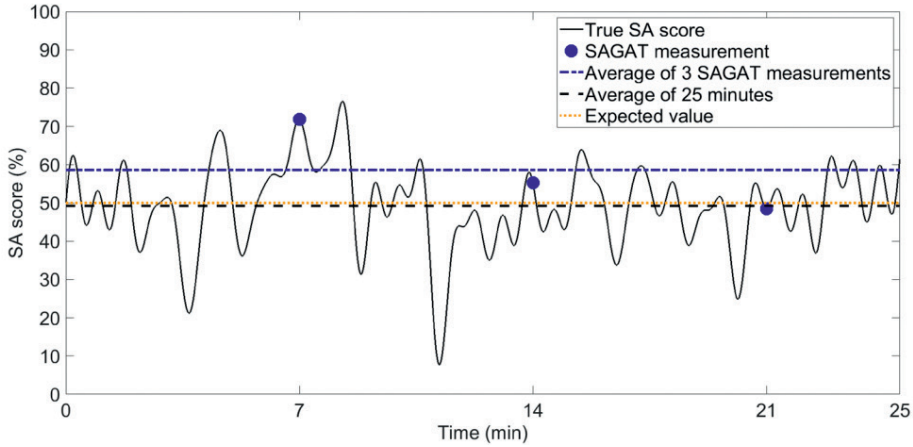


Figure. 1 Hypothetical illustration of a human's true SA score during a 25 min task. Three simulation freezes were assumed during which the SAGAT score was probed (at 7, 14, and 21 min). Here, we assumed that SA varies continuously, which is plausible, given that the state of technological systems (velocity, mass flow, etc.) is necessarily continuous due to laws of physics. However, SA could also change in discrete steps because the system state may manifest in discrete forms (e.g., warning lights) and because perception may resemble discrete steps also (as illustrated with multistable perception; Leopold and Logothetis 1999)

Fourth, the SAGAT requires the participant to bring aspects of the task environment forward into conscious attention and to answer corresponding queries. However, what an operator reports in a query does not necessarily reflect his/her knowledge of the situation. According to dualprocessing theories, which distinguish between unconscious (i.e., implicit, automatic) and conscious (i.e., explicit, controlled) processes (Evans 2003; Kahneman 2011; Kihlstrom 2008), it is the unconscious processes that are evoked based on situational triggers. Reflexes and instincts are the most basic examples of non-conscious behaviors in response to environmental stimuli. Implicit cognitive processes may also be acquired through practice. For example, after sufficient practice, drivers perform certain elementary tasks, such as changing gears, without overt conscious attention (Shinar et al. 1998, see also Morgan and Hancock 2008). Other familiar paradigms, such as the Stroop task, provide a further illustration that participants process the meaning of stimuli unconsciously, whether they want to or not. Endsley (1995) acknowledged that "data may be processed in a highly automated fashion and

thus not be in the subject's awareness" (p. 72). However, she argued that the intrusion of unconscious processes represents only a small threat to SAGAT, by invoking three lines of reasoning. First, she argued that participants who fill out a SAGAT response sheet are able to extract situational content from long-term memory despite the fact that information has been processed automatically. Second, she reasoned that the multiple-choice response style of SAGAT facilitates access from memory, as opposed to when being asked openended questions. The third argument was that participants are likely aware that they will complete a SAGAT query, which in turn enhances memorization and recall. Whether these assertions are true, and whether the recognition associated with the third argument does not interfere with memory capacity in the first place, requires further research. In sum, from the preceding observations, it would appear that the individual responds to environments often founded upon information not readily available to conscious introspection.

The fifth issue with SAGAT is that it measures SA intermittently rather than continuously, and therefore, it does not capture the dynamics of SA (Stanton et al. 2015). According to the law of large numbers, when administering the SAGAT on a small number of instances, one obtains a relatively imprecise estimate of the long -run expected value (Fig. 1). Moreover, when sampling at a limited rate, one does not capture higher frequencies in the signal. It is the fluctuations in SA that can be valuable sources of information for assessing cause-and-effect relationships regarding how changes of the environment, interoperator communication, or task feedback influence SA.

Finally, the SAGAT task-freeze approach fails to take account of the situated cognition phenomenon (Stanton et al. 2015). People rely on artifacts to hold information on their behalf (Hutchins 1995; Sparrow et al. 2011). A study by Walker et al. (2009) comparing the communication modes of voice-only (i.e., no video, no data), video, and data-link in a distributed planning task showed that the SAGAT method could lead to the decision to use voice only. This was due to the fact that as the communication media became richer the SAGAT scores became poorer. As Stanton et al. (2015) reported, "The explanation lies in that the greater the support from the environment, the less the person has to remember as the artifacts in the system hold the information" (p. 46). It seems a falsehood to divorce cognition from context. Similarly, Chiappe et al. (2015) argued that SAGAT is an inappropriate method to measure SA as blanking the screens prevents operators "from accessing externally represented information that they are used to obtaining in this way when engaged in a task" (p. 40).

3. TOWARDS SA ESTIMATION FROM EYE MOVEMENTS IN RELATION TO THE TASK ENVIRONMENT

We have indicated that it would be of considerable value to be able to assess SA in realtime. Here, we select eye movements as a candidate variable for the dynamic measurement of SA. The use of eye movement counteracts each of the above limitations of the SAGAT, as eye movement measurements are available on a continuous basis, can be obtained without interrupting or disturbing the ongoing task, do not require the operator to bring task elements to explicit memory, and are, therefore, free from issues of memory decay.

Humans rotate their eyes to orient the high-resolution fovea to the part of their scene that promises to render the greatest information. According to the eye-mind hypothesis, gaze direction is a strong correlate of cognitive activity (Just and Carpenter 1980; Yarbus 1967). Furthermore, according to the thesis of situated cognition, cognitive activity routinely exploits structure in the natural and social environment (Robbins and Aydede 2009). Given such an assumption, it should be feasible to identify some aspects of SA from eye-movements in relation to the task environment.

First, we illustrate the potential of eye movements through the lens of driving, which is a common task with strong safety implications (World Health Organization 2015). Driving is predominantly a visual task (Sivak 1996; Van der Horst 2004). In a review of more than half a century of driving safety research, Lee (2008) concluded that most crashes occur because "drivers fail to look at the right thing at the right time" (p. 525). Car driving involves much more than mere object detection, as drivers look ahead (i.e., 'preview') to anticipate and respond to what will happen next (e.g., Deng et al. 2016; Donges 1978). Research on how drivers extract relevant information from the task environment has often been reported under the heading of 'hazard perception' or 'hazard anticipation', which are terms now often equated with SA (Underwood et al. 2013; Horswill and McKenna 2004).

Recent research in this area has indicated that hazard precursors are discriminative between inexperienced and experienced drivers (Garay-Vega and Fisher 2005; Underwoodet al. 2011). Precursors are visual cues that place critical demands on the driver's understanding and projection of an unfolding situation (cf. Levels 2 and 3 SA), such as the example shown in Fig. 2. Drivers with high SA are expected to be more likely to glance at the sports car (Level 1 SA), because the state of the sports car is informative about future collision risks (Levels 2 and 3 SA). Thus, in order to compute a driver's SA, an algorithm first has to establish critical features in the environment (e.g., a sports car is inching out), and whether the driver has attended to this feature. To clarify, a lot of eye movements in an environment with many task-relevant objects may signal high SA (because the driver scans these task-relevant objects), whereas the same eye

movements in an environment with a small number of critical objects may signal low SA (i.e., the driver is distracted).

Figure 2. A precursor used in previous SA research. Participants watch an unfolding scene. "This moped rider is about to pass a sports car with a driver in it and the front wheels turned to the left. If this sports car pulls out, the moped rider has to brake or swerve to the left. Has the participant driver noticed the sports car?" (from Vlakveld 2011)

4. AN EMPIRICAL DEMONSTRATION OF MEASURING SA BY MEANS OF EYE MOVEMENTS IN RELATION TO THE TASK ENVIRONMENT

Here, we provide a demonstration by means of experimental results as to how SA can be extracted from eye movements in relation to task conditions. The results herein are based on an experiment presented in Eisma et al. (2018).

We used a visual sampling paradigm in which participants viewed a series of moving dials (Senders 1983). The participant's head was fixed via a head support (i.e., no postural changes). Thus, the human rotated the eyes to perceive the status of the display. Even though the task was chosen to be simple, it encapsulates the essential monitoring features of supervisory control of a dynamic system. This paradigm has its origins in a study by Fitts et al. (1950), which has been called "the first major Human Factors study" (Senders 2016).

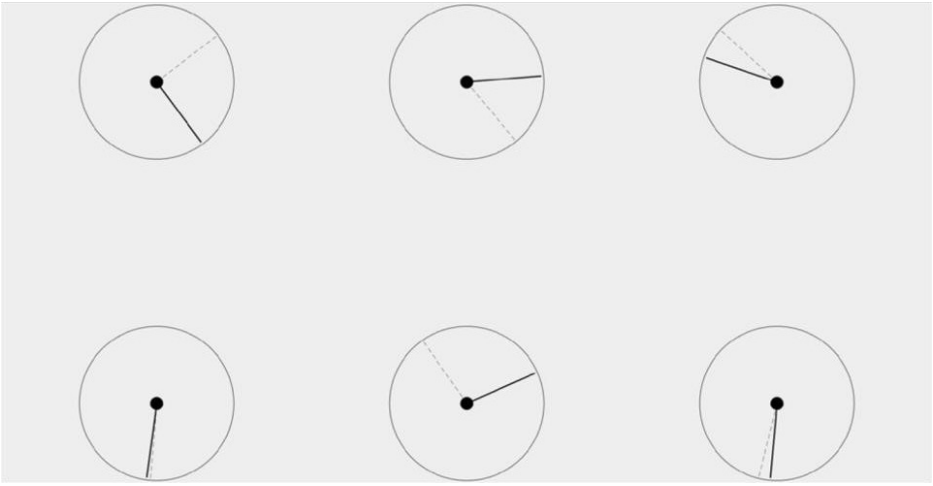


Figure 3. Screenshot of one of the seven videos. The dashed line is the threshold. The solid line is the pointer

We express the amount of 'grip' on the environment as the percentage of resemblance between observed and ideal conditions, where 100% means optimal performance, and a low or zero percentage means that the operator's mind is wandering or the operator is asleep or unconscious, being completely disengaged or oblivious to the task. Accordingly, we define a 'sampling score' that defines how well the human observer has scanned the status of the dynamic displays.

4.1 Experimental methods

4.1.1 Participants

Participants were 86 university students (21 female, 65 male) with a mean age of 23.44 years (SD = 1.52) (Eisma et al. 2018). The original sample consisted of 91 participants, but data for five participants proved invalid due to computer faults, eye-tracker limitations, or data storage errors. The research was approved by the Ethics Committee of the TU Delft under the title 'Update of Visual Sampling Behavior and Performance with Changing Information Bandwidth'. All participants provided written informed consent.

4.1.2 Experimental tasks

Participants viewed seven 90-s videos on a 24-inch monitor having a resolution of 1920 × 1080 pixels. An EyeLink 1000 Plus was used to track the participants' eye movements. Each video showed six circular dials with moving pointers (as in Senders 1983). The pointer movement was a random signal with a bandwidth that differed between the six dials (0.03, 0.05, 0.12, 0.20, 0.32, and 0.48 Hz; as in Senders 1983). The threshold (dashed line, see Fig. 3) was a random angle that differed for each of the 42 dials (7 videos × 6 dials). In each of the seven videos, the pointer signals had a mean of 0 deg

(i.e., relative to where the threshold was defined) and a standard deviation of 50.1 deg. The signal realization was different for each of the 42 dials, and the bandwidth ordering per dial was different for the seven videos. Each participant viewed the same seven videos in randomized order. An example video is provided in the supplementary materials.

4.1.3 Experimental procedures

Participants first completed a training of 20 s during which a single dial was shown. Participants were instructed to press the spacebar when a pointer crossed the threshold from either direction. The screen blanked after each video, and participants immediately completed a paper and pencil test about the current (Question 1), past (Question 2), and future (Question 3) states of the pointers (Fig. 4).

4.1.4 Dependent measures

First, we calculated a performance score per participant. This score was defined as the percentage of threshold crossings for which the participant pressed the spacebar. In total, there were between 74 and 115 threshold crossings per video. Per crossing, a 'hit' was counted if the participant pressed the spacebar within 0.5 s (i.e., between −0.5 and +0.5 s) of the moment of the crossing (Eisma et al. 2018). A spacebar press could not be assigned to more than one threshold crossing, and no more than one hit could be assigned to a threshold crossing.

Second, we calculated a visual sampling score per participant. This measure of SA was defined as the percentage of threshold crossings for which the participant fixated on a 420 × 420 pixel square surrounding the dial, within 0.5 s of the moment of the threshold crossing.

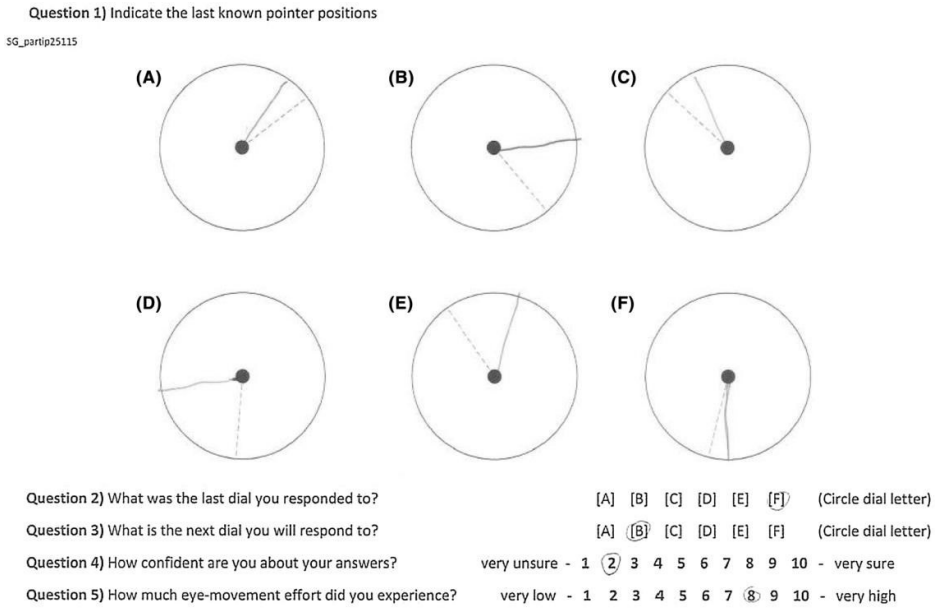


Figure 4. The form completed by a participant (using a blue pencil) after one of the seven videos. In Question (1) participants drew a line, while in Questions (2–5) they circled an answer

Third, we calculated a freeze-probe score for each participant. This score was defined as the percentage of 42 dials for which the participant drew a line on the correct side of the threshold.[11] The correct side meant that the line drawn by the participant occurred within the same clockwise or counterclockwise angular direction (i.e., from the threshold at 0° to ±180°) as the 'ground truth' (i.e., the pointer position at the end of the video) . If a participant did not draw a line (which happened in six out of 3588 dials) the score for this particular dial was marked as incorrect. We chose this binary definition (correct vs. incorrect side from the threshold) of the freeze probe score because alternative measures (e.g., absolute difference between the drawn angle and the threshold angle) may be prone to bias. More specifically, we observed that participants tended to draw the line near the threshold (if they were uncertain); this approach would yield a low error score (because the pointer indeed moves around the threshold) even when the participant was merely guessing. Furthermore, a binary scoring corresponds with the SAGAT, where participants have to tick a response which can be either correct or incorrect.

1 We used image recognition in MATLAB to extract where participants had drawn the line. Participants used a blue pen, which could be relatively easily differentiated from the black/white background. The image recognition was found to have a mean accuracy of 0.14° (determined from the threshold which was printed on paper versus the known location of the threshold).

For three of 86 participants, freeze-probe data were unavailable in one to two out of seven forms. Furthermore, for three other participants, due to computer/calibration issues, eye-tracking data for one to three out of seven videos were unavailable. These participants were retained in the analysis, using only relevant and acceptable data.

4.2 Experimental results

Participants viewing behavior was found to strongly relate to the state and dynamics of the dials. With high replication correspondence to the results of Senders (1983), glance frequency, dwell time, and dwell time per glance were evidenced as a function of task signal bandwidth (for details, see Eisma et al. 2018).

Table 1 shows a crosstabulation of the sampling and performance score per threshold crossings. It can be seen that if a dial was not visually sampled in the right 1-s time frame (i.e., surrounding when a pointer crossed a threshold), then it was unlikely (28.4%) that the participant pressed the spacebar in that same 1-s time frame. Conversely, if a dial was sampled, then the participant pressed the spacebar in more than 50% (60.8%) of the threshold crossings. The phi coefficient (equivalent to the Pearson product-moment correlation coefficient) between the visual sampling score and the performance score equaled 0.31. The correlation between the visual sampling score and the performance score at the level of participants was 0.78 (see Fig. 5, right).

Table 1. Cross-tabulation of the number of times a dial was (not) sampled and a spacebar was (not) pressed, for each threshold crossing

	Dial not sampled	Dial sampled
Spacebar not pressed	13,135 (71.6%)	13,445 (39.2%)
Spacebar pressed	5208 (28.4%)	20,839 (60.8%)
Total	18,343 (100%)	34,284 (100%)

The average freeze-probe score among participants was 57.7% (SD = 8.6%), which is slightly better than the expected value of 50% if participants were simply guessing. Participants had little confidence in their answers (Question 4 in Fig. 4): The average score was 4.08 (SD = 1.50) on the scale from 1 (very unsure) to 10 (very sure) (Fig. 4). Participants' freeze-probe score exhibited a moderate correlation with their performance score, $r = 0.20$ (Fig. 5, left).

The mean score on Question 2 (last dial) was 31.3% (SD = 18.6%) with respect to the last threshold crossing, and 29.6% (SD = 16.0%) with respect to the last space bar 'hit', whereas the mean score on Question 3 (next dial the participant will respond to) was 17.1% (SD = 13.6%), where 16.7% would be expected based on guessing alone. The scores on Questions 2 and 3 did not correlate significantly with the visual sampling score or freeze-probe score (all *r*s between −0.10 and 0.13).

In summary, we have shown that there is a moderate correlation (r = 0.31) between visual sampling and task performance at the level of threshold crossings, and a strong correlation at the level of participants (r = 0.78). Furthermore, it appears that participants had difficulty memorizing the state of the dials even though they filled out the form immediately after completing the task. In other words, how people sampled the dials was more strongly predictive of performance than what they memorized about the dials.

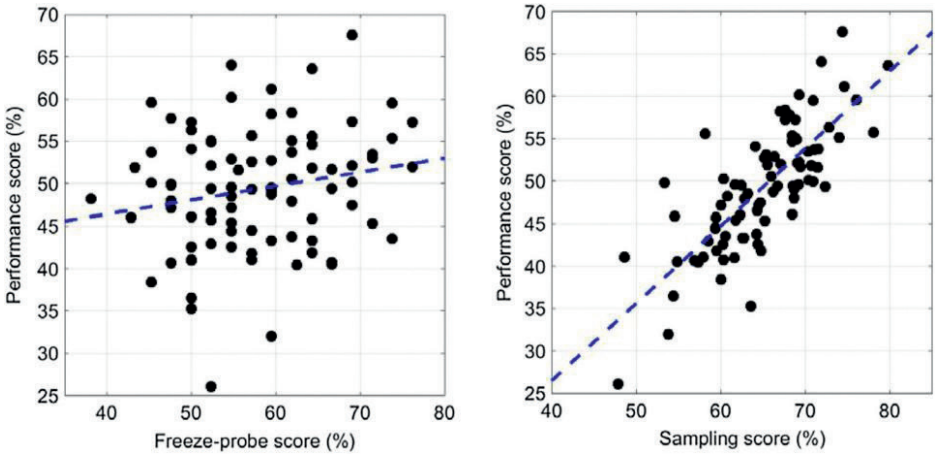


Figure. 5 The association between freeze-probe score and performance score (left panel, r = 0.20), and the association between visual sampling score and performance score (right panel, r = 0.78). Each marker represents a participant. The dashed line is a linear least-squares fit

5 DISCUSSION

5.1 Main findings

This paper aimed to outline several fundamental limitations of SAGAT and examine whether an eye-based measure of SA can be more predictive of task performance than a freeze-probe method. We argued that the SAGAT has the following limitations: (1) time delays between the freeze moment and the moment of answering the queries, (2) task interruption/disruption, (3) a disconnect from the ongoing task, (4) the need to bring the situation to conscious memory, (5) intermittent rather than continuous SA measurement, and (6) a failure to take situated cognition into account. Such fundamental limitations can help account for contentious empirical results regarding the validity of the SAGAT found in the literature (as reviewed in Sect. 1.2).

Building upon earlier work by Moore and Gugerty (2010), we have here shown that task performance can be predicted through eye-tracking measurements in relation to the state of the task environment in a more accurate manner than achieved by SAGAT. More specifically, correlations between visual sampling scores and performance scores were

0.31 at the level of threshold crossings and 0.78 at the level of individuals. In contrast, freeze-probe scores were low and showed weak associations with task performance. These results may be insufficiently compelling for real-time feedback applications, as the number of false positives and misses were rather high. However, we note that these calculations are binary (the timing or likelihood of glances were not considered), and therefore, there are multiple opportunities for improvement in both the sensitivity and specificity.

5.2 Hardware and software requirements

What hardware and software would be needed to implement a realtime SA assessment method based on eye movements in real-life situations? If the present approach were to be implemented in car driving, for example, high-end cameras would be needed that capture eye movements regardless of vibrations, lighting conditions, and driver's headgear such as caps, eyeglasses, and sunglasses. In the 1980s, physiological measurement tools were often bulky with limited capabilities (see Moray and Rotenberg 1989, for a study on human-automation interaction with gaze analyses at only 2 Hz). Consistent with Moore's (1965) law, however, computers have become considerably smaller and faster, and it is perhaps only a matter of time until we have the availability of ubiquitous eye-tracking cameras.

Additionally, the state of the environment has to be known. The ground truth could be human-generated as in SAGAT (choosing what to measure from the eyes and the task environment) or it could be computergenerated (e.g., using algorithms to determine what are relevant objects to look at). The latter approach requires databases (e.g., maps), sensors (e.g., cameras, radar), and analysis methods (e.g., instance segmentation of camera images). These capabilities are already being developed, for example for autonomous driving applications (Uhrig et al. 2016). A computer-generated ground truth should be able to establish that the turning of the sports car wheels shown in Fig. 2 is a hazard precursor, and that a situationally aware driver can be expected to have had their eyes towards this cue. Other operators (e.g., road users) may be part of the environment and so their states and dynamics should also be inputs for the model. Wickens et al. (2003, 2008) previously introduced a computational model of attention and SA based on the prior works of Senders. In their model, the probability of attending to an area is a weighted average of not only bandwidth as in Senders (1964, 1983), but also saliency (i.e., the conspicuity of information), effort (i.e., the visual angle between areas, where a larger angle is expected to inhibit scanning), and value (i.e., the importance of tasks served by the attended event). Attention to an area (i.e., Level 1 SA) is used to update human understanding of the current and future state of the system. This model appears to be a useful point of departure for developing a comprehensive algorithm for real-time SA assessment.

In real-life situations, multiple bodily signals (e.g., posture, see Riener et al. 2008) may need to be considered simultaneously as an input to a computational model, in order to infer SA. For example, it may be hard to extract SA related to strategies with long time constants from eye movements only. Additionally, the eye-mind hypothesis does not hold in a strong sense. In driving, a sizeable portion of collisions are caused by the looked-but-failed-to-see phenomenon, as well as related phenomena such as staring, mind wandering, and inattentional blindness (Herslund and Jørgensen 2003; White and Caird 2010). In other words, although the driver is fixated on a relevant stimulus, attention may covertly reside elsewhere. More research then appears to be needed to examine the validity of eye-based SA in complex supervisory tasks. In particular, it needs to be examined how eye-based SA can be employed in teams, especially in situations where different human actors and cognitive artifacts have conflicting information or intentions, and where task knowledge needs to be communicated between those agents (e.g., Salmon et al. 2008; Stanton et al. 2017; Vanderhaegen and Carsten 2017).

In sum, real-time SA assessment in outdoor environments is an engineering challenge, but not an unrealistic one considering the ongoing developments in sensors and artificial intelligence. So framed, our method is not fundamentally different from SAGAT, as both incorporate a comparison with ground truth. The difference is that SAGAT responses are explicitly reported by participants and cannot be extracted from veridical situations but only from simulated ones. In our case, the ground truth concerned the moments of threshold crossings of the pointer, whereas Moore and Gugerty (2010) defined specific aircraft as “important” within their air traffic control task environment upon which to evaluate the SA (estimation) construct. We recommend that researchers move beyond the use of paper and pencil tests of SA, and address and embrace the above developments to achieve the goal of ubiquitous SA assessment.

5.3 Differences from performance measurements and operator state assessments

Our proposal differs from performance-based measures of SA (Durso and Gronlund 1999; Gutzwiller and Clegg 2013; Prince et al. 2007; Sarter and Woods 1995). Performance-based SA suffers from circular reasoning, in the sense that it defines SA in terms of performance, but performance is what SA should prospectively predict in the first place (see Warm et al. 2008 recognizing the same paradox when mental resources are defined as task performance). Furthermore, in real-life tasks, such as supervision of highly automated systems, continuous performance measurements are often simply unavailable because the operator provides input only occasionally. In the present experiment, we asked participants to press the space bar when the pointers exceeded a threshold value. In reality, humans are often passive supervisors without an active performance task or overt responses to record.

Our approach also differs from operator-state assessment systems in general. For example, in driving, several sensor technologies exist that detect whether a driver is fatigued or distracted (Barr et al. 2009; Blanco et al. 2009; Dong et al. 2011). Such systems may make use of measures of head movement, blink rate, eyelid closure, or gaze direction in any and all combinations and then provide feedback according to a multivariate algorithm (optionally combined with physiological and performance measures). The problem is that many of these systems measure the operator's behaviors without considering the environmental context in which behavior is embedded, and so may attack the issue of awareness per se, but do not reflect *situation* awareness specifically.

5.4 Future prospects

Hoffman and Hancock (2014) lamented that in many Human Factors investigations that are aimed at investigating why participants behave the way they do, researchers apparently never "bothered to ask the participants any questions after the experiment was over." Thus, there is clearly an inherent value in self-report and freeze-probe techniques for measuring SA, but we regard our approach to be in the long-term more promising and valuable for engineering applications that rely on real-time SA assessment, such as training and adaptive automation. Finally, we believe that the shortcomings of SAGAT, such as its reliance on memory skills and its disruption, also apply to many other SA procedures. For example, online probe measures, such as the situation awareness rating technique (SART), may be even more disruptive than SAGAT, but are likely less susceptible to issues of memory decay. As Salmon et al. (2006) noted, the SAGAT is "by far the most commonly used approach, and also the technique with the most associated validation evidence" (p. 228). Thus, it appears to be fair that we featured SAGAT as a target to which a new SA measure should be compared.

We have provided a demonstration as to how predictive-valid SA can be computed from eye movements and task features alone. From an engineering viewpoint, the human can be viewed as a machine (albeit a machine made of living tissue) and therefore all of a human's behavior has to have physical causes. The more accurate and information-rich the eye-movement and environment measurements become, the more opportunities arise for observing SA from these measurements. Concomitantly, the need for invoking indirect measures such as SAGAT then diminishes.

5.5 Limitations of the present experiment

The present task, in which participants had to watch a number of dials, may be regarded as arbitrary and unrepresentative of complex real-life situations such as control rooms and cockpits. However, our supervisory control task was intentionally designed to be abstract to provide a generic account of SA measurement. Moreover, our task replicated previous research of Senders (1983) and resembles the seminal work of Fitts et al. (1950),

wherein pilots monitored a number of flight instruments (e.g., airspeed, directional gyro, engine instruments, altitude, vertical speed). We argue that our sampling task captures the essence of supervisory control—an area that Sheridan (1980) forecasted as increasingly relevant—in that operators have to monitor automation/ instruments and detect anomalies (i.e., threshold crossings).

It may also be argued that our present freeze-probe measurement does not capture whether operators understand the situation (Level 2 SA) and anticipate what will happen (Level 3 SA). However, a review of the SAGAT shows that it is often used in simple tasks and includes simple items, such as items where participants have to recall the location of aircraft or cars (Endsley 2000b). That is, it seems that the use of our freeze-probe method does not fundamentally differ from the use of a typical SAGAT.

Participants performed poorly on the freeze-probe task and had little confidence in their answers. It is plausible that participants would score higher on freeze-probe queries if the supervisory task were interactive and meaningful (e.g., operating a nuclear power plant). As explained by Durso and Gronlund (1999), operators apply several strategies to reduce demands on working memory. Such strategies include focusing on the important information only, chunking of meaningful information, and restructuring the environment. Although our supervisory task did not allow for such strategies, our results do illustrate that participants were hardly able to remember the situation they had seen a few seconds before, a finding that is consistent with the notion that operators process information unconsciously (for explanation see Sect. 2). Eye-tracking seems a viable tool for measuring whether/when an operator has looked at specific objects (e.g., aircraft, cars), and provides a more direct indicator of SA than self-reported recall of the presence of objects or system states. Future research should establish whether SA based on eye movements in relation to the task environment can predict future, as opposed to concurrent performance, whether the criterion validity upholds in semantically rich tasks with longer time constants and correlated signals, and whether real-time feedback/control provided based on SA can enhance safety and productivity in operational settings.

Another limitation of the present study is that the participants were students at a technical university. As shown by Wai et al. (2009), engineering students score highly on intelligence-related tests, including tests of spatial ability. Accordingly, it is likely that engineering students have higher working memory capacity and would score better on the freeze-probe task than the general population. Because freeze-probe scores would likely be even lower in a sample that is representative of the entire population, our postulations and results against freeze-probe SA measurements are conservatively drawn. Another limitation of using engineering students is restriction of range (Hunter et al. 2006). That is, because of the relatively homogenous sample, correlations between

task-performance scores, visual sampling scores, and freeze-probe scores are likely attenuated as compared to correlations in a sample with a broad range of abilities. The issue of range restriction is especially pertinent for SA research, which is often concerned with specific groups of experts, such as pilots, military personnel, or air traffic control operators (Durso and Gronlund 1999).

6. CONCLUSIONS

It is concluded that the SAGAT suffers from time delays, task disruption, a disconnect from the ongoing task, a bias towards conscious recall, intermittent measurement, and a lack of measuring the situatedness of SA. We advanced a method to circumvent these limitations by calculating SA based on eye movements in relation to the task environment. We conclude that real-time SA based on eyes in relation to the task environment is moderately correlated with performance at the event level and strongly correlated with task performance at the level of individual participants.

REFERENCES

Barr L, Popkin S, Howarth H, Carroll RJ (2009) An evaluation of emerging driver fatigue detection measures and technologies: final report (report no. FMCSA-RRR-09-005). Volpe National Transportation Systems Center, Cambridge

Blanco M, Bocanegra JL, Morgan JF, Fitch GM, Medina A, Olson RL et al (2009) Assessment of a drowsy driver warning system for heavy-vehicle drivers: final report (report no. DOT 811). Virginia Tech Transportation Institute, Blacksburg, p 117

Bogossian F, Cooper S, Cant R, Beauchamp A, Porter J, Kain V et al (2014) Undergraduate nursing students' performance in recognising and responding to sudden patient deterioration in high psychological fidelity simulated environments: an Australian multi-centre study. Nurse Educ Today 34:691–696

Carsten O, Vanderhaegen F (2015) Situation awareness: valid or fallacious? Cogn Technol Work 17:157–158

Chiappe D, Strybel TZ, Vu KPL (2015) A situated approach to the understanding of dynamic situations. J Cogn Eng Decis Mak 9:33–43

Cummings ML, Guerlain S (2007) Developing operator capacity estimates for supervisory control of autonomous vehicles. Hum Factors 49:1–15

Dekker SW (2015) The danger of losing situation awareness. Cogn Technol Work 17:159–161

Deng T, Yang K, Li Y (2016) Where does the driver look? Top-down-based saliency detection in a traffic driving environment. IEEE Trans Intell Transp Syst 17:2051–2062

Dong Y, Hu Z, Uchimura K, Murayama N (2011) Driver inattention monitoring system for intelligent vehicles: a review. IEEE Trans Intell Transp Syst 12:596–614

Donges E (1978) A two-level model of driver steering behavior. Hum Factors 20:691–701

Durso FT, Gronlund SD (1999) Situation awareness. In: Durso FT, Nickerson RS, Schvaneveldt RW, Dumais ST, Lindsay DS, Chi MTH (eds) Handbook of applied cognition. Wiley, New York, pp 283–314

Durso FT, Hackworth CA, Truitt TR, Crutchfield J, Nikolic D, Manning CA (1998) Situation awareness as a predictor of perfor-mance for en route air traffic controllers. Air Traffic Control Q 6:1–20

Durso FT, Bleckley MK, Dattel AR (2006) Does situation awareness add to the validity of cognitive tests? Hum Factors 48:721–733

Eisma YB, Cabrall CDD, De Winter JCF (2018) Visual sampling processes revisited: replicating and extending Senders (1983) using modern eye-tracking equipment. IEEE Trans Hum Mach Syst. https://doi.org/10.1109/THMS.2018.2806200

Endsley MR (1988) Design and evaluation for situation awareness enhancement. Proc Hum Factors Ergon Soc Ann Meet 32:97–101

Endsley MR (1995) Measurement of situation awareness in dynamic systems. Hum Factors 37:65–84

Endsley MR (2000a) Theoretical underpinnings of situation awareness: a critical review. In: Endsley M, Garland DJ (eds) Situation awareness: analysis and measurement. Lawrence Erlbaum Associates, Mahwah, pp 3–32

Endsley MR (2000b) Direct measurement of situation awareness: validity and use of SAGAT. In: Endsley MR, Garland DJ (eds) Situation awareness analysis and measurement. Lawrence Erlbaum Associates, Mahwah, pp 147–174

Endsley MR (2015a) Situation awareness misconceptions and misunderstandings. J Cogn Eng Decis Mak 9:4–32

Endsley MR (2015b) Final reflections situation awareness models and measures. J Cogn Eng Decis Mak 9:101–111

Evans JSB (2003) In two minds: dual-process accounts of reasoning. Trends Cogn Sci 7:454–459

Fitts PM, Jones RE, Milton JL (1950) Eye movements of aircraft pilots during instrument-landing approaches. Aeronaut Eng Rev 9:24–29

Flach JM (1995) Situation awareness: proceed with caution. Hum Factors 37:149–157

Garay-Vega L, Fisher DL (2005) Can novice drivers recognize foreshadowing risks as easily as experienced drivers? Proceedings of the third international driving symposium on human factors in driver assessment, training and vehicle design. Rockport, ME

Gardner AK, Kosemund M, Martinez J (2017) Examining the feasibil-ity and predictive validity of the SAGAT tool to assess situational awareness among surgical trainees. Simul Healthc J Soc Simul Healthc 12:17–21

Gugerty LJ (1997) Situation awareness during driving: explicit and implicit knowledge in dynamic spatial memory. J Exp Psychol Appl 3:42–66

Gugerty LJ (1998) Evidence from a partial report task for forgetting in dynamic spatial memory. Hum Factors 40:498–508

Gugerty LJ, Tirre WC (2000) Individual differences in situation aware-ness. In: Endsley MR, Garland DJ (eds) Situational awareness analysis and measurement. Erlbaum, Mahwah, pp 249–276

Gutzwiller RS, Clegg BA (2013) The role of working memory in levels of situation awareness. J Cogn Eng Decis Mak 7:141–154

Hancock PA (2014) Automation: how much is too much? Ergonomics 57:449–454

Herslund M-B, Jørgensen NO (2003) Looked-but-failed-to-see-errors in traffic. Accid Anal Prev 35:885–891

Hoffman RR, Hancock PA (2014) Words matter. Hum Factors Ergon Soc Bull 57:3–7

Hogan MP, Pace DE, Hapgood J, Boone DC (2006) Use of human patient simulation and the situation awareness global assessment technique in practical trauma skills assessment. J Trauma Acute Care Surg 61:1047–1052

Horswill MS, McKenna FP (2004) Drivers' hazard perception ability: situation awareness on the road. In: Banbury S, Tremblay S (eds) A cognitive approach to situation awareness. Ashgate, Aldershot, pp 155–175

Hunter JE, Schmidt FL, Le H (2006) Implications of direct and indirect range restriction for meta-analysis methods and findings. J Appl Psychol 91:594–612

Hutchins E (1995) How a cockpit remembers its speeds. Cogn Sci 19:265–288

Ikuma LH, Harvey C, Taylor CF, Handal C (2014) A guide for assessing control room operator performance using speed and accuracy, perceived workload, situation awareness, and eye tracking. J Loss Prev Process Ind 32:454–465

Just MA, Carpenter PA (1980) A theory of reading: from eye fixation to comprehension. Psychol Rev 87:329–354

Kahneman D (2011) Thinking, fast and slow. Farrar, Straus and Giroux, New York

Kihlstrom JF (2008) The automaticity juggernaut. In: Baer J, Kaufman JC, Baumeister RF (eds) Are we free? Psychology and free will. Oxford University Press, New York, pp 155–180

Lee JD (2008) 50 years of driving safety research. Hum Factors 50:521–528

Leopold DA, Logothetis NK (1999) Multistable phenomena: changing views in perception. Trends Cogn Sci 3:254–264

Lo JC, Sehic E, Brookhuis KA, Meijer SA (2016) Explicit or implicit situation awareness? Measuring the situation awareness of train traffic controllers. Transp Res Part F 43:325–338

Loft S, Bowden V, Braithwaite J, Morrell DB, Huf S, Durso FT (2015) Situation awareness measures for simulated submarine track management. Hum Factors 57:298–310

McGowan A, Banbury S (2004) Interruption and reorientation effects of a situation awareness probe on driving hazard anticipation. Proc Hum Factors Ergon Soc Ann Meet 48:290–294

Merleau-Ponty M (1945) Phenomenology of perception. Routledge, Abingdon

Millot P (2015) Situation awareness: Is the glass half empty or half full? Cogn Technol Work 17:169–177

Moore GE (1965) Cramming more components onto integrated circuits. Electronics 38:114–117

Moore K, Gugerty L (2010) Development of a novel measure of situation awareness: the case for eye movement analysis. Proc Hum Factors Ergon Soc Ann Meet 54:1650–1654

Moray N, Rotenberg I (1989) Fault management in process control: eye movements and action. Ergonomics 32:1319–1342

Morgan JF, Hancock PA (2008) Estimations in driving. In: Castro C (ed) Human factors of visual and cognitive performance in driving. Taylor and Francis/CRC, Boca Raton, pp 51–62

Neisser U (1976) Cognition and reality: principles and implications of cognitive psychology. W. H. Freeman/Times Books/Henry Holt & Co, New York

O'Brien KS, O'Hare D (2007) Situational awareness ability and cognitive skills training in a complex real-world task. Ergonomics 50:1064–1091

Parasuraman R, Sheridan TB, Wickens CD (2008) Situation awareness, mental workload, and trust in automation: viable, empirically supported cognitive engineering constructs. J Cogn Eng Decis Mak 2:140–160

Pew RW (1994) Situation awareness: the buzzword of the 90s. Gate-way 5:1–4

Pierce RS, Strybel TZ, Vu KPL (2008) Comparing situation awareness measurement techniques in a low fidelity air traffic control simulation. Proceedings of the 26th international congress of the aeronautical sciences (ICAS). Anchorage, AS

Prince C, Ellis E, Brannick MT, Salas E (2007) Measurement of team situation awareness in low experience level aviators. Int J Aviat Psychol 17:41–57

Riener A, Ferscha A, Matscheko M (2008) Intelligent vehicle handling: steering and body postures while cornering. In: Brinkschule U, Ungerer T, Hochberger C, Spallek RG (eds) Architecture of com-puting systems–ARCS 2008. Springer, Berlin, pp 68–81

Robbins P, Aydede M (eds) (2009) The Cambridge handbook of situated cognition. Cambridge University Press, Cambridge

Rushton JP, Brainerd CJ, Pressley M (1983) Behavioral development and construct validity: the principle of aggregation. Psychol Bull 94:18–38

Salmon P, Stanton N, Walker G, Green D (2006) Situation awareness measurement: a review of applicability for C4i environments. Appl Ergon 37:225–238

Salmon PM, Stanton NA, Walker GH, Baber C, Jenkins DP, McMaster R, Young MS (2008) What really is going on? Review of situation awareness models for individuals and teams. Theor Issues Ergon Sci 9:297–323

Salmon PM, Stanton NA, Walker GH, Jenkins D, Ladva D, Rafferty L, Young M (2009) Measuring situation awareness in complex systems: comparison of measures study. Int J Ind Ergon 39:490–500

Sarter NB, Woods DD (1995) How in the world did we ever get into that mode? Mode error and awareness in supervisory control. Hum Factors 37:5–19

Sechrest L (1963) Incremental validity: a recommendation. Educ Psychol Meas 23:153–158

Senders JW (1964) The human operator as a monitor and controller of multidegree of freedom systems. IEEE Trans Hum Factors Electron 1:2–5

Senders JW (1983) Visual sampling processes (Doctoral dissertation). Katholieke Hogeschool Tilburg, the Netherlands

Senders JW (2016) History of human factors [video file]. https://www.youtube.com/watch?v=QZH97Xy6NA8&t=2801s. Accessed 1 Feb 2018

Sheridan TB (1980) Computer control and human alienation. Technol Rev 83:60–73

Shinar D, Meir M, Ben-Shoham I (1998) How automatic is manual gear shifting? Hum Factors 40:647–654

Sivak M (1996) The information that drivers use: is it indeed 90% visual? Perception 25:1081–1089

Smith K, Hancock PA (1995) Situation awareness is adaptive, externally directed consciousness. Hum Factors 37:137–148

Sohn YW, Doane SM (2004) Memory processes of flight situation awareness: Interactive roles of working memory capacity, long-term working memory, and expertise. Hum Factors 46:461–475

Sparrow B, Liu J, Wegner DM (2011) Google effects on memory: cognitive consequences of having information at our fingertips. Science 333:776–778

Stanton NA, Salmon PM, Walker GH (2015) Let the reader decide: a paradigm shift for situation awareness in socio-technical systems. J Cogn Eng Decis Mak 9:44–50

Stanton NA, Salmon PM, Walker GH, Salas E, Hancock PA (2017) State-of-science: situation awareness in individuals, teams and systems. Ergonomics 60:449–466

Strybel TZ, Vu KPL, Kraft J, Minakata K (2008) Assessing the situation awareness of pilots engaged in self spacing. Proc Hum Factors Ergon Soc Ann Meet 52:11–15

Uhrig J, Cordts M, Franke U, Brox T (2016) Pixel-level encoding and depth layering for instance-level semantic labeling. arXiv:1604.05096

Underwood G, Crundall D, Chapman P (2011) Driving simulator validation with hazard perception. Transp Res Part F 14:435–446

Underwood G, Ngai A, Underwood J (2013) Driving experience and situation awareness in hazard detection. Saf Sci 56:29–35

Van der Horst R (2004) Occlusion as a measure for visual workload: an overview of TNO occlusion research in car driving. Appl Ergon 35:189–196

Van der Maas HL, Dolan CV, Grasman RP, Wicherts JM, Huizenga HM, Raijmakers ME (2006) A dynamical model of general intelligence: the positive manifold of intelligence by mutualism. Psychol Rev 113:842–861

Vanderhaegen F, Carsten O (2017) Can dissonance engineering improve risk analysis of human-machine systems? Cogn Technol Work 19:1–12

Vlakveld WP (2011) Hazard anticipation of young novice drivers (Doctoral dissertation). University of Groningen. SWOV dissertation series. SWOV Institute for Road Safety Research, Leidschendam, The Netherlands

Wai J, Lubinski D, Benbow CP (2009) Spatial ability for STEM domains: aligning over 50 years of cumulative psychological knowledge solidifies its importance. J Educ Psychol 101:817–835

Walker GH, Stanton NA, Salmon P, Jenkins D (2009) How can we support the commander's involvement in the planning process? An exploratory study into remote and co-located command planning. Int J Ind Ergon 39:456–464

Warm JS, Parasuraman R, Matthews G (2008) Vigilance requires hard mental work and is stressful. Hum Factors 50:433–441

White CB, Caird JK (2010) The blind date: the effects of change blindness, passenger conversation and gender on looked-but-failed-to-see (LBFTS) errors. Accid Anal Prev 42:1822–1830

Wickens CD (2008) Situation awareness: review of Mica Endsley's 1995 articles on situation awareness theory and measurement. Hum Factors 50:397–403

Wickens CD (2015) Situation awareness its applications value and its fuzzy dichotomies. J Cogn Eng Decis Mak 9:90–94

Wickens C, McCarley J, Thomas L (2003) Attention-situation awareness (A-SA) model. In: Foyle DC, Goodman A, Hooey BL (eds) Proceedings of the 2003 conference on human performance modeling of approach and landing with augmented displays (NASA/ CP-2003-212267). NASA, Moffett Field, pp 189–225

Wickens CD, McCarley JS, Alexander AL, Thomas LC, Ambinder M, Zheng S (2008) Attention-situation awareness (A-SA) model of pilot error. In: Foyle DC, Hooey BL (eds) Human performance modeling in aviation. Taylor and Francis/CRC, Boca Raton, pp 213–242

World Health Organization (2015) WHO global status report on road safety 2015. World Health Organization, Geneva

Yarbus AJ (1967) Eye movements and vision. Plenum Press, New York

CHAPTER 5

Attention Distribution While Detecting Conflicts between Converging Objects: An Eye-Tracking Study

Eisma, Y. B., Looijestijn, A. E., & De Winter, J. C. F. (2020). Attention distribution while detecting conflicts between converging objects: An eye-tracking study. *Vision, 4,* 34.

ABSTRACT

In many domains, including air traffic control, observers have to detect conflicts between moving objects. However, it is unclear what the effect of conflict angle is on observers' conflict detection performance. In addition, it has been speculated that observers use specific viewing techniques while performing a conflict detection task, but evidence for this is lacking. In this study, participants (N = 35) observed two converging objects while their eyes were recorded. They were tasked to continuously indicate whether a conflict between the two objects was present. Independent variables were conflict angle (30, 100, 150 deg), update rate (discrete, continuous), and conflict occurrence. Results showed that 30 deg conflict angles yielded the best performance, and 100 deg conflict angles the worst. For 30 deg conflict angles, participants applied smooth pursuit while attending to the objects. In comparison, for 100 and especially 150 deg conflict angles, participants showed a high fixation rate and glances towards the conflict point. Finally, the continuous update rate was found to yield shorter fixation durations and better performance than the discrete update rate. In conclusion, shallow conflict angles yield the best performance, an effect that can be explained using basic perceptual heuristics, such as the 'closer is first' strategy. Displays should provide continuous rather than discrete update rates.

1. INTRODUCTION

In many types of occupations and daily activities, humans have to make decisions concerning spatial events that involve moving objects. A large number of empirical studies exist on this topic, for example in the area of car driving. These studies usually apply an egocentric perspective, where the observer performs temporal judgments while moving relative to one or more vehicles in the environment (e.g., [1,2]).

A less studied type of spatial task concerns the detection of a conflict between allocentric objects that move towards each other. This type of task has mainly been studied in air traffic control and other aviation contexts (e.g., [3,4]). As early as 1947, Gibson [5] described a variety of motion-picture-based tests for training and selection of air force personnel. One such test concerned the depiction of two animated planes, one overtaking the other. Before the overtaking point was reached, the planes disappeared behind a cloud, and the operator had to indicate at which imagined point the two planes would collide. Besides aviation, allocentric conflict detection tasks occur in areas such as gaming (e.g., robot combat or classical multidirectional shooter games; [6]) and monitoring of mobile agents. In the future, human operators may have to supervise the safe separation of drones [7], teleoperated cars [8], or mobile robots [9].

Human performance in allocentric conflict detection or arrival-time judgment tasks has also been studied in its own right without reference to a specific application (e.g., [10,11]). Kimball [12] argued that “time predictions about future positions of moving objects are made many times a day by virtually everyone” (p. 935). Much of the upcoming literature review pertains to air traffic control tasks, but we do not mean to imply that the application of our research is constrained to air traffic control.

1.1. The Effect of Conflict Angle on Conflict Detection Performance

Several studies have examined the accuracy with which operators detect whether two given aircraft are in conflict. Studies among university students [13] and licensed and trainee air traffic controllers [3] have found that participants are more likely to intervene when presented with a smaller conflict angle (45, 90, and 135/150 deg were used). These two studies also showed that participants frequently made false alarms, especially with smaller conflict angles. That is, with small conflict angles in particular, participants often indicated that there was a conflict when in fact there was a large minimum separation between the two aircraft.

Loft et al. [3] argued that air traffic controllers work under constraints of uncertainty: they have to estimate the aircraft trajectories and set criteria regarding whether to intervene. They noted that the effects of uncertainty are higher at smaller conflict angles because for small conflict angles, a small position estimation error can result in a large overlap of trajectories (see also [14], as cited in [15]). However, a potential

confounder is that Loft et al. [3] defined a conflict as a loss of separation of 5 nautical miles, not a collision of two aircraft. Accordingly, a smaller conflict angle implies a longer period of violation of separation.

Pompanon and Raufaste [16] found, in a study among 556 novices who applied for a flight school, that 90 deg conflict angles resulted in shorter response times and higher accuracy in estimating the intersection point compared to smaller (45 deg) and higher (135 deg) conflict angles. These results appear to confirm the findings of Loft et al. [3] in that the small conflict angle of 45 deg yielded a large uncertainty/dispersion of the estimated intersection point. In a study among university students, Law et al. [11] found that moving objects on a parallel convergent (180 deg) trajectory yielded a higher conflict detection accuracy as compared to an oblique (45 deg) and perpendicular (90 deg) trajectory. They reported that "the effect of configuration seems to be primarily associated with visual scanning. As the objects are presented farther apart, accuracy decreases" [11] (p. 1188). However, eye movements were not examined in that study.

In summary, based on the above, it is unclear which type of conflict angle yields the best conflict detection performance, because results are contingent on various assumptions. Adding to the complication, in previous research, operators had to provide a 'conflict' or 'no conflict' response as quickly as possible, after which further responses were no longer possible [13,17]. This approach, in which only one data point per trial is obtained, cannot provide full insight into how operators accumulate evidence, or how they adjust their perceptual-cognitive strategies, as time elapses.

In a review on conflict detection, Xu and Rantanen [18] argued that operators might use various visual-cognitive processes during conflict detection tasks. For example, if the operator knows that the speeds of the convergent aircraft are equal, then the operator merely has to detect whether the distances of the two aircraft towards the conflict point are equal to infer that a conflict will occur. An alternative process would be to visually or cognitively extrapolate the motion of the aircraft [18]. Another model was proposed by Neal and Kwantes [13]. Their model assumes that operators iteratively sample evidence regarding the state of the world and accumulate it over time. They used their model to predict response times in a conflict decision task for different conflict angles but offered no further validations.

The visual-cognitive processes mentioned above seem plausible, but a weakness of the reviewed research is that the processes were not observed, but only inferred from performance measures. As pointed out by Xu and Rantanen [18], "the detection accuracy and the response time examined in the previous investigations seem to be the measures of the final product of conflict detection" (p. 3), not the actual process. Accordingly, the researchers recommended further research into operators' conflict detection processes.

1.2. The Potential of Eye-Tracking in Conflict Detection Research

Eye-tracking can be used to unravel the relationships between the geometry of a scenario containing converging objects and operators' visual-cognitive information processes. According to the strong eye-mind hypothesis [19], the location of observers' eye fixations coincides with what the observer is mentally processing at that moment.

Thus far, only a few studies have examined how observers distribute their visual attention during allocentric conflict detection tasks. One relevant study is by Hunter and Parush [20], who recorded eye movements of university students observing two aircraft on a convergent trajectory. They found that the participants were more likely to scan between the two aircraft than towards the collision point. Based on this finding, they argued that "attention to the collision site may not be as essential to conflict detection as was previously thought" (p. 1732). However, important limitations are that Hunter and Parush's [20] research was conducted with a relatively inaccurate head-mounted eye-tracker and that participants were presented with only one scenario. Furthermore, their analyses did not provide insight into how eye movement measures varied as the scenario progressed.

Another relevant study using eye-tracking was conducted by Pompanon and Raufaste [21]. In this work, 30 experienced air traffic controllers were asked to detect conflicts between two aircraft that flew on conflicting or divergent trajectories and at the same or different altitudes. Based on recorded first glances to areas of interest as well as response times, the authors proposed a model of human information processing. In short, this model asserted that operators first assess whether the aircraft converge or diverge. Next, they assess altitude differences between the two aircraft, and then they try to recognize geometric patterns in the trajectories and deduce whether the aircraft are in conflict. Pompanon and Raufaste's [21] work is a good example of the usefulness of eye-tracking for this type of research. However, similar to Hunter and Parush [20], they did not show how the eye movements changed over time.

1.3. Study Aims

This study aimed to examine the effect of conflict angle on operators' performance in allocentric conflict detection tasks. The above literature suggests that conflicts involving small conflict angles are easiest to detect yet prone to false positives. However, these results can be explained by the size of the separation zone (typically 5 nautical miles) and not by conflict angle per se. Another limitation of the existing research is that in the majority of the studies, participants provided only a single response per trial. Various studies have forwarded hypotheses of the visual-cognitive strategies that observers use while performing allocentric conflict detection tasks. However, the use of such strategies cannot be validly derived from response times alone.

In our experiment, we varied conflict angles from small (30 deg) to intermediate (100 deg) and large (150 deg) and examined how observers distribute their attention between two moving objects on a convergent trajectory. Measurements of eye movements and conflict detection were made continuously during each trial. Because the literature provides no clear leads, we formulated no a-priori hypotheses regarding the effect of conflict angle. In addition to eye movements, we acquired measures of conflict detection performance and self-reported difficulty. These two complementary measures were thought to reflect the difficulty of the conflict detection task.

In this study, we offered an additional manipulation: stimulus update rate. That is, all stimuli were offered with continuous movements and with discrete movements. Current radar systems provide discrete information because the radar sweeps at a fixed rate. A literature review by Chen and Thropp [22] of 50 empirical studies about the effect of update rate (i.e., frame rate) showed that a reduction of update rate is associated with a decrease of task performance. Therefore, we expected that performance in the conflict detection task would be better if the converging objects moved in a continuous as compared to a discrete manner. In Chen and Thropp's [22] literature review, performance reductions were found in a variety of tasks, including placement, tracking, target recognition, and perceptual judgment tasks. For target recognition and perceptual judgments tasks, however, low update rates were sometimes found to yield a performance equivalent to baseline [22]. For example, a driving simulator study by Van Erp and Padmos [23] found no significant effect of update rate (3 up to 30 Hz were tested) on speed estimation accuracy. Accordingly, conflict detection performance may be unaffected by update rate.

2. METHODS

2.1. Participants

Thirty-six persons participated in the experiment. They were students or recently graduated persons at the Delft University of Technology. The data of one participant were excluded because this participant did not perform the task as instructed. The remaining 35 participants consisted of 19 males and 16 females, between 18 and 31 years old (mean = 22.8, standard deviation (*SD*) = 2.91). Participants were offered compensation of 5 Euro for their time. This research was approved by the University's Human Research Ethics Committee. A written informed consent form was signed by all participants before the start of the experiment.

2.2. Participants' Task

Participants watched a total of 36 videos, each containing a scenario of 20 s. In each scenario, two dots were linearly moving towards each other (Figure 1). Participants were instructed to keep the spacebar pressed when they thought the dots would collide.

Figure 1. Screenshots of one scenario at three moments. **Left**: Beginning of the video, **Middle**: 10 s into the video, **Right**: 15 s into the video. This is a non-conflict scenario, with a conflict angle of 150 deg.

After each scenario, participants indicated to what extent they agreed with the statement: "The task was difficult" on a scale from 0 (completely disagree) to 10 (completely agree). Next, the participant was shown his/her performance score for that scenario. The performance score was computed as the percentage of time that the spacebar was correctly pressed or released, depending on whether the scenario contained a conflict or no conflict, respectively.

Before the experiment, a calibration of the eye tracker was performed. Furthermore, participants were familiarized with the task using one training scenario with discrete stimuli. This scenario had a different geometry from the scenarios of the experiment. In the training scenario, a collision was presented. A break of a few minutes was held halfway during the experiment. The experiment lasted about 30 min per participant.

2.3. Apparatus

Eye movements were recorded at 2000 Hz using the SR-Research Eyelink 1000 Plus. Participants were asked to place their head in the head support. The stimuli were displayed on a 24 inch BENQ monitor with a resolution of 1920 × 1080 pixels (531 × 298 mm). The refresh rate of the monitor was 60 Hz. Based on an approximate distance of 91 cm between the monitor and the participant's eyes, the monitor subtended viewing angles of 33 deg horizontally and 19 deg vertically.

2.4. Independent Variables

The first independent variable is the conflict angle between the two dots. In the literature, conflict angles have been divided into three categories: 0–60 deg (overtake), 60–120 deg (crossing), and 120–180 deg (head-on) [24]. For this experiment, one angle from each of these categories was used, namely 30, 100, and 150 deg.

The second independent variable was the update rate consisting of two levels: discrete and continuous. For the continuous stimuli, the update rate of the location of the dots was set equal to the video frame rate (30 frames per second). For the discrete stimuli, the update rate of the location of the dots was 2 times per second.

The third independent variable was the conflict outcome. In real-life tasks, objects have to retain a safe separation. For example, in air traffic control, aircraft have to be

separated at least five nautical miles from each other (e.g., [25]). In this experiment, no separation zone was defined around the dots. The dots could either collide or not collide.

Each combination of independent variables was repeated three times in a different configuration, which meant that we rotated the entire stimulus with 0, 45, and 90 deg. In summary, participants were presented with 36 scenarios (3 conflict angles × 2 stimulus update rates × 2 conflict outcomes × 3 configurations). The sequence of the 36 scenarios was randomized for each participant.

2.5. Design of the Stimuli

The scenario consisted of a white (RGB: 0.9 0.9 0.9) background of 1920 × 960 pixels, on which two circular dots with a diameter of 18 pixels were shown (RGB: 0.1 0.1 0.1). The two dots ('aircraft') were moving at constant speeds and the same altitude on straight, converging courses [18].

Herein, we expressed the dimensions of the scenarios in pixels, as this information allows for exact reproduction of our methods. For our setup, a distance of 100 pixels on the screen corresponds to an angular range of approximately 1.7 deg. The speed of both dots was 26.4 pixels/s (528 pixels in 20 s or about 0.45 deg/s) during the entire experiment. For the discrete stimuli, the dots jumped forward 13.2 pixels per frame. This distance amounts to a change in visual angle of about 0.2 deg, which means that participants could keep a jump of a dot within foveal vision without re-fixating.

Dot 1 always started 480 pixels from the center of the screen. Dot 1 moved through the middle and ended 48 pixels from the mid-point. The heading of Dot 2 was determined by the conflict angle (i.e., 30, 100, or 150 deg) relative to Dot 1. For scenarios in which the dots collided, Dot 2 started 480 pixels from the center of the screen and ended 48 pixels from the midpoint, just as Dot 1. Thus, the collision occurred 18.3 s into the 20 s scenario.

For non-conflict scenarios, Dot 2 started with a 58-pixel offset so that the closest point of approach with respect to Dot 1 was 58 pixels, occurring 18.3 s into the scenario. This closest point of approach was determined using pilot tests. We ensured that the conflict detection task was not too easy (which would be when participants could easily see that no conflict would occur, e.g., at the beginning of the scenario) and not too difficult (i.e., which would be when participants could distinguish conflict from no conflict only during the last few seconds of the scenario).

Table 1. Characteristics of Scenarios 1–18.

Scenario Number	Conflict Angle (deg)	Dot 1 Heading (deg)	Dot 2 Heading (deg)	Dot 1 Coordinate (*x*, *y* in pixels)		Dot 2 Start Coordinate (*x*, *y* in pixels)		Conflict	Relative Distance to CP at Start (Dot 1/Dot 2)	Dot 2 Passing Dot 1
1	**30**	270	300	1440	480	1375	723	Yes	1	
2	30	225	195	1299	141	1081	15	Yes	1	
3	30	180	150	960	0	717	65	Yes	1	
4	30	270	240	1440	480	1431	225	No	0.89	Behind
5	30	225	195	1299	141	1113	−34	No	0.89	Behind
6	30	180	150	960	0	705	9	No	0.89	Behind
7	100	270	10	1440	480	874	955	Yes	1	
8	100	225	125	1299	141	563	205	Yes	1	
9	100	180	280	960	0	1435	566	Yes	1	
10	100	270	10	1440	480	840	909	No	1.08	In front
11	100	225	325	1299	141	1178	868	No	1.08	In front
12	100	180	280	960	0	1389	600	No	1.08	In front
13	150	270	120	1440	480	541	239	Yes	1	
14	150	225	75	1299	141	493	606	Yes	1	
15	150	180	330	960	0	1201	899	Yes	1	
16	150	270	60	1440	480	529	664	No	1.03	In front
17	150	225	75	1299	141	468	554	No	0.97	Behind
18	150	180	330	960	0	1144	911	No	1.03	In front

Note. Scenarios 1–18 were presented with continuous movements of the dots. Scenarios 19–36 are identical to Scenarios 1–18, but with discrete movement. The (0, 0) coordinate is the left top corner of the video. Heading angles of 0, 90, 180, and 270 deg are north, east, south, and west, respectively. CP = conflict point.

All dimensions, including the closest point of approach of 58 pixels, were dimensionless. Participants were not provided with any reference about a numeric distance of speed and were therefore unable to interpret the task in reference to particular standards for safe separation.

An overview of the scenarios is shown in Table 1. Scenarios 19–36 are identical to Scenarios 1–18, but with discrete instead of continuous stimuli.

2.6. Dependent Variables

A median filter with a 100 ms interval was used to smoothen the raw eye-tracking data. When no eye data were available (e.g., during a blink), linear interpolation was used. The dependent variables were defined as follows:

Performance score (%): The performance score was computed as the percentage of time the participant had the spacebar correctly pressed or not pressed. For example, if a participant held the spacebar pressed between 7 and 12 s during a non-conflict scenario, the performance score for that participant in that scenario was (20 − 5 s)/20 s·100% = 75%.

Self-reported difficulty (0–10): A difficulty score between 0 and 10 was provided by the participants after each scenario, on a scale from (completely disagree) to 10 (completely agree).

Fixation rate (Hz): A higher fixation rate means that participants sample more elements from the scenario per time unit. For calculating the fixation rate, the eye-tracking data were partitioned into saccades and fixations in the same way as in Eisma, Cabrall, and De Winter [26]. First, the gaze speed was filtered with a Savitzky-Golay filter with order 2 and a frame length of 41. A saccade velocity threshold of 2000 pixels per second was used. The minimum fixation duration was set at 40 ms.

Mean fixation duration (s): During fixations, participants acquire information from the visual array. This measure is inversely related to the fixation rate. A longer mean fixation duration means that participants focused longer on the same element of the scenario.

Mean saccade amplitude (pixels): Saccade amplitude is another common measure in eye-tracking research [27]. A higher mean saccade amplitude indicates that participants have a broader spread of fixations.

Mean fixation amplitude (pixels): Smooth pursuit is a type of eye movement that involves the continuous movement of the eyes while tracking a moving object. From a visual inspection of participants' *x* and *y* gaze coordinates, it became apparent that some fixations contained smooth pursuit, where participants followed one of the two dots. According to Holmqvist et al. [28], smooth pursuit is not easily identified, and "it is currently an open research problem to develop a robust and generic algorithm for

such a purpose" (p. 152). Holmqvist et al. [28] also explained that standard velocity algorithms typically assign smooth pursuit data in the same category as fixations. Indeed, we observed that some fixations had a large amplitude, that is, the eyes traveled on the screen but without rapid saccade. Herein, we used the following measure of the degree of smooth pursuit: "as with saccades, the amplitude of smooth pursuit can also be calculated as the shortest distance between the points of on- and off-set" (p. 319). Holmqvist et al. [28] explained that this measure only works well when the direction of pursuit remains relatively constant, which we believe is a valid assumption in our case because the dots moved linearly. In summary, for each fixation, the straight-line distance from the start to the end moment of the fixation was computed and used as an index of the amount of pursuit. Thus, we did not classify fixations into smooth pursuit and no smooth pursuit but calculated the amplitude for each fixation.

Gaze coordinates on area of interest (AOI) (%): In accordance with Hunter and Parush [20], we assessed whether participants focused on the conflict point or one of the two dots. More specifically, we calculated the percentage of time that participants' gaze coordinates were on one of the two dots within a radius of 100 pixels, hereafter referred to as the 'dots AOI'. Additionally, we calculated the percentage of time that participants' gaze coordinates were on the conflict point within a radius of 100 pixels, hereafter referred to as the 'CP AOI'. In the case of non-conflict scenarios, the conflict point was defined as the mean of the coordinates of Dots 1 and 2 at their closest point of approach.

2.7. Statistical Analyses

First, scores on the dependent variables were compared between the scenarios with continuous and discrete update rates. Because we wanted to assess the main effect of update rate, paired-samples *t*-tests were used. Additionally, a three-way repeated-measures analysis of variance (ANOVA) of the fixation rate was performed to examine the effects of update rate (continuous versus discrete), conflict angle (30, 100, 150 deg), and conflict occurrence (no conflict versus conflict). Based on the small interaction effects with update rate (continuous versus discrete), we decided to aggregate the results of the continuous and discrete stimuli in subsequent analyses. Differences between the three conflict angles were compared using a repeated-measures ANOVA. Pairs of conflict angles were compared using paired-samples *t*-tests. *p*-values smaller than 0.05 were considered significant. Effect sizes between conditions were expressed as Cohen's *d* and Cohen's d_z. Cohen's d_z describes the within-subjects effect size [29].

3. RESULTS

3.1. Continuous Versus Discrete Stimuli

Table 2 shows that participants had a significantly higher performance score for continuous stimuli as compared to discrete ones. Furthermore, participants had a lower fixation rate and higher mean fixation duration for discrete stimuli as compared to the continuous stimuli. The effects for mean saccade amplitude, mean fixation amplitude, and self-reported difficulty were not statistically significant between continuous and discrete stimuli.

Table 2. Means and standard deviations (*SD*) of the dependent variables for continuous and discrete stimuli, as well as results of paired *t*-tests between the scores for continuous and discrete stimuli.

	Continuous Stimuli		**Discrete Stimuli**					
	Mean	***SD***	**Mean**	***SD***	***t*(34)**	***p***	**Cohen's *d***	**Cohen's d_z**
Fixation rate (Hz)	1.145	0.298	0.993	0.280	6.66	<0.001	0.53	1.13
Mean fixation duration (ms)	813	235	905	269	−4.40	<0.001	−0.36	−0.74
Mean saccade amplitude (pixels)	182	31	179	33	1.47	0.151	0.10	0.25
Mean fixation amplitude (pixels)	36	12	35	13	0.83	0.411	0.08	0.14
Performance score (%)	70.8	5.59	68.3	5.56	2.09	0.044	0.46	0.35
Self-reported difficulty (0–10)	5.30	1.28	5.43	1.24	–1.31	0.198	−0.11	−0.22

Note. The results for each participant were averaged for the 18 continuous scenarios and the 18 discrete scenarios.

A three-way repeated measures full-factorial ANOVA for the fixation rate showed a significant effect of update rate ($F(1,34) = 44.3$, $p < 0.001$, $\eta_p^2 = 0.57$), conflict angle ($F(2,68) = 267.7$, $p < 0.001$, $\eta_p^2 = 0.89$), and conflict occurrence ($F(1,34) = 18.0$, $p < 0.001$, $\eta_p^2 = 0.35$).

The interaction effect for update rate × conflict angle was small but significant ($F(2,68) = 5.35$, $p = 0.007$, $\eta_p^2 = 0.14$). Paired *t*-tests were conducted to examine the effect of update rate per conflict angle. For conflict trials, the effect of update rate increased with increasing conflict angle: $t(34) = 3.42$, 4.13, and 5.07, and $p = 0.002$, $p < 0.001$, and $p < 0.001$, for conflict angles of 30, 100, and 150 deg, respectively. A similar trend was observed for non-conflict trials: $t(34) = 2.81$, 4.93, and 4.01, and $p = 0.008$, $p < 0.001$, and $p < 0.001$, for conflict angles of 30, 100, and 15 deg, respectively. This interaction effect may be due to the fact that larger conflict angles involved a higher number of fixations (see Figure 2). The interaction effect for update rate × conflict occurrence was

not significant ($F(1,34) = 0.05$, $p = 0.825$, $\eta_p^2 = 0.00$). Because the interaction effects with update rate were small, we averaged the results for the continuous and discrete stimuli in subsequent analyses.

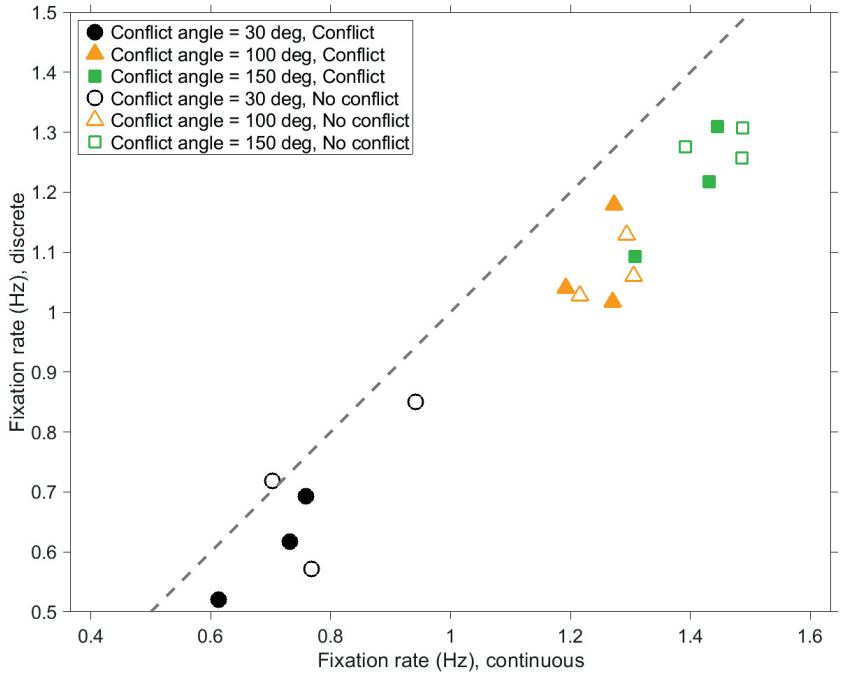


Figure 2. Mean number of fixations per second for scenarios with discrete stimuli versus scenarios with continuous stimuli. Each marker represents the average of 35 participants. The dashed line is the line of equality.

3.2. Effect of Conflict Angle on Conflict Detection Performance

Conflict scenarios: An evaluation of the spacebar pressings shows that conflicts were detected earlier for 30 deg conflict angles as compared to 100 and 150 deg conflict angles (Figure 3). For example, 5 s into the scenario, 36% of participants had pressed the spacebar in 30 deg scenarios, compared to 20% and 19% of participants in 100 and 150 deg scenarios, respectively. A repeated-measures ANOVA of the performance scores also showed a significant difference between conflict angles, $F(2,68) = 12.2$, $p < 0.001$. Paired t-tests showed significant differences between 30 and 100 deg scenarios ($t(34) = 5.05$, $p < 0.001$, $d = 0.61$, $d_z = 0.85$), between 30 and 150 deg scenarios ($t(34) = 3.60$, $p = 0.001$, $d = 0.54$, $d_z = 0.61$), but not between 100 and 150 deg scenarios ($t(34) = -0.92$, $p = 0.364$, $d = -0.12$, $d_z = -0.16$).

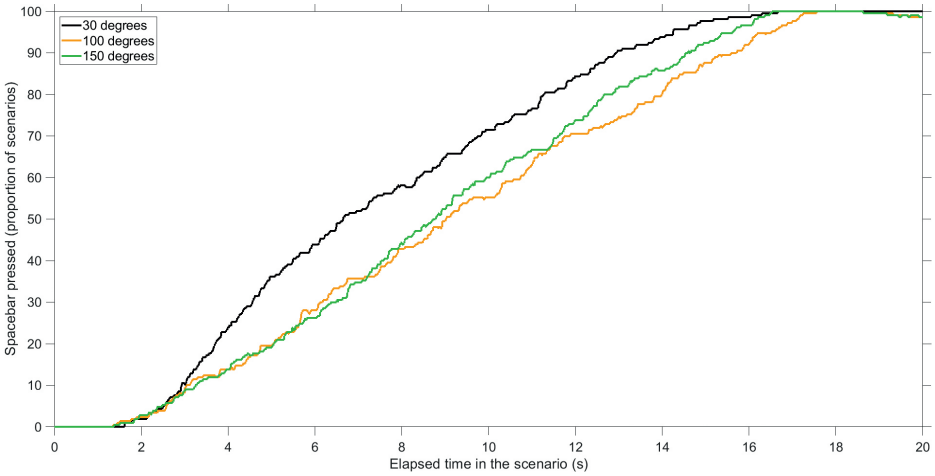


Figure 3. Percentage of participants who pressed the spacebar at that point in time during the scenario, for conflict scenarios. The proportion is calculated for 210 scenarios (35 participants × 6 scenarios per conflict angle).

Non-conflict scenarios: Furthermore, with 100 deg conflict angles, many participants falsely believed that there would be a conflict (Figure 4). The percentage of participants who falsely reported a conflict at a particular moment during the scenario was maximally 30%, 56%, and 40% for 30 deg (at 9.61 s), 100 deg (at 12.20 s), and 150 deg (at 10.07 s) conflict angles, respectively (see Figure 4 for a visualization). A repeated-measures ANOVA of the performance scores showed a significant difference between conflict angles, $F(2,68) = 10.5$, $p < 0.001$. Paired *t*-tests showed significant differences between 30 and 100 deg scenarios ($t(34) = 5.17$, $p < 0.001$, $d = 0.97$, $d_z = 0.87$), between 100 and 150 deg scenarios ($t(34) = -3.21$, $p = 0.003$, $d = -0.62$, $d_z = -0.54$), but not between 30 and 150 deg scenarios ($t(34) = 1.11$, $p = 0.274$, $d = 0.27$, $d_z = 0.19$).

Conflict and non-conflict scenarios combined: Figure 5 shows that 30 deg conflict angles yielded the highest performance, and 100 deg conflict angles the lowest. A repeated-measures ANOVA of the performance scores showed a significant difference in performance between conflict angles, $F(2,68) = 25.3$, $p < 0.001$, $\eta_p^2 = 0.43$. Paired *t*-tests showed significant differences between 30 and 100 deg scenarios ($t(34) = 8.21$, $p < 0.001$, $d = 1.61$, $d_z = 1.39$), between 30 and 150 deg scenarios ($t(34) = 3.41$, $p = 0.002$, $d = 0.79$, $d_z = 0.58$), and between 100 and 150 deg scenarios ($t(34) = -3.29$, $p = 0.002$, $d = -0.72$, $d_z = -0.56$).

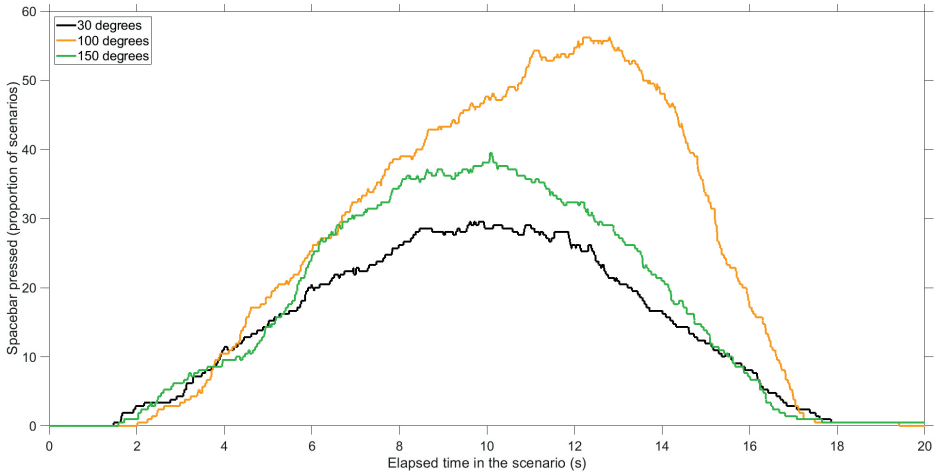


Figure 4. Percentage of participants who pressed the spacebar at that point in time during the scenario, for non-conflict scenarios. The proportion is calculated for 210 scenarios (35 participants × 6 scenarios per conflict angle).

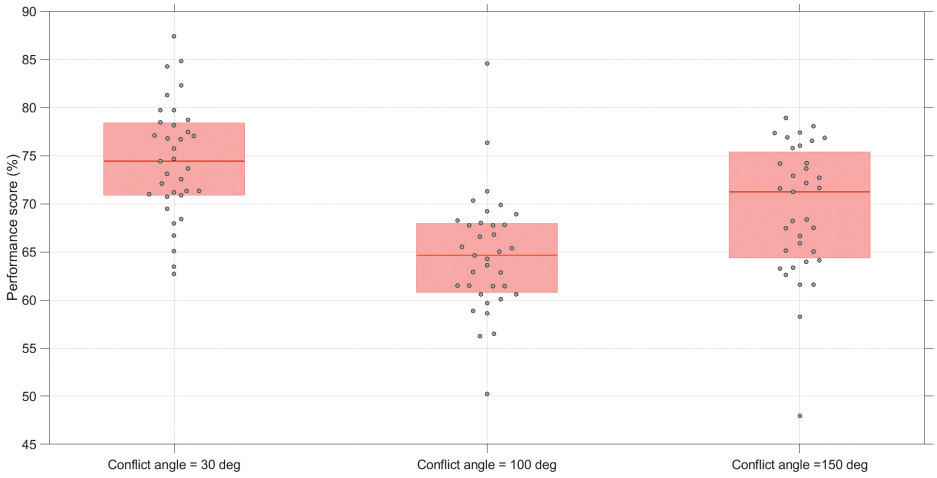


Figure 5. Boxplots of the performance scores per conflict angle. The score for each participant represents the average of 12 scenarios (conflict scenarios and non-conflict scenarios combined).

3.3. Effect of Conflict Angle on Self-Reported Difficulty (Conflict and Non-Conflict Scenarios Combined)

The self-reported difficulty was higher for the 100 deg conflict angle as compared to the other two conflict angles (Figure 6). A repeated-measures ANOVA showed significant differences between the three angles, $F(2,68) = 30.1$, $p < 0.001$, $\eta_p^2 = 0.47$. Paired t-tests further showed significant differences between 30 and 100 deg ($t(34) = -7.61$, $p < 0.001$, $d = -0.77$, $d_z = -1.29$), between 30 and 150 deg ($t(34) = -3.42$, $p = 0.002$, $d = -0.38$, $d_z = -0.58$), and between 100 and 150 deg ($t(34) = 4.71$, $p < 0.001$, $d = 0.37$, $d_z = 0.80$).

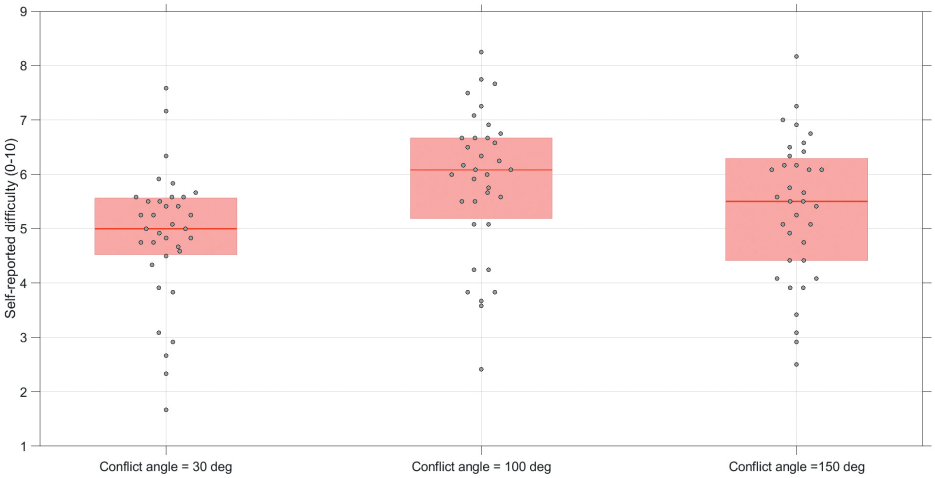


Figure 6. Boxplots of the self-reported difficulty scores per conflict angle. The score for each participant represents the average of 12 scenarios (conflict scenarios and non-conflict scenarios combined).

3.4. Effect of Conflict Angle on Eye Movements (Conflict and Non-Conflict Scenarios Combined)

Videos showing the gaze coordinates for all scenarios are available in the online data archive. From an inspection of the videos, we noted that participants predominantly looked at the dots (dots AOI) or in between the dots (i.e., close to an imaginary line connecting the two dots). Figure 7 (top) provides a video snapshot, illustrating that the participants sampled in between the dots or directly at the dots. However, in the 100 deg and 150 deg scenarios, participants sometimes directed their gaze towards the conflict point angles (see Figure 7, bottom, for an illustration for looking towards the conflict point).

Figure 8 provides further information about participants' looking behavior at AOIs as a function of elapsed time in the scenario. It can be seen that for 30 deg conflict angles, participants predominantly looked at the dots AOI and hardly looked at the conflict point (CP AOI). For 100 deg conflict angles, and especially for 150 deg conflict angles, participants did look at the conflict point to some extent. Most of the remaining time was spent looking in between the dots (see also Figure 7).

As shown in Figure 9, mean fixation durations were longer for smaller conflict angles. A repeated-measures ANOVA showed significant effects of conflict angle, $F(2,68) = 68.7$, $p < 0.001$, $\eta_p^2 = 0.72$. Paired t-tests also showed significant differences in fixation duration between 30 and 100 deg ($t(34) = 8.27$, $p < 0.001$, $d = 1.05$, $d_z = 1.40$), between 30 and 150 deg ($t(34) = 11.10$, $p < 0.001$, $d = 1.46$, $d_z = 1.88$), and between 100 and 150 deg ($t(34) = 4.76$, $p < 0.001$, $d = 0.41$, $d_z = 0.80$).

Further analysis of the data revealed dynamic viewing patterns as a function of elapsed time during the scenario. The saccade amplitude showed interpretable patterns: saccades had a larger amplitude earlier in the scenario as well as for larger conflict angles (Figure 10). This decrease of amplitude can be explained by the fact that the distance between the dots linearly decreases with elapsed time. At 18.3 s in the scenario, the two dots collided. When the outcome of the scenario (i.e., collision or no collision) becomes evident, participants sometimes sample elsewhere on the screen, which can explain the increase of saccade amplitude near the end of the scenario. The fixation amplitude describes whether participants tracked an object using pursuit movement. The fixation amplitude also increased near the end of the scenario, especially for the small conflict angle of 30 deg (Figure 11).

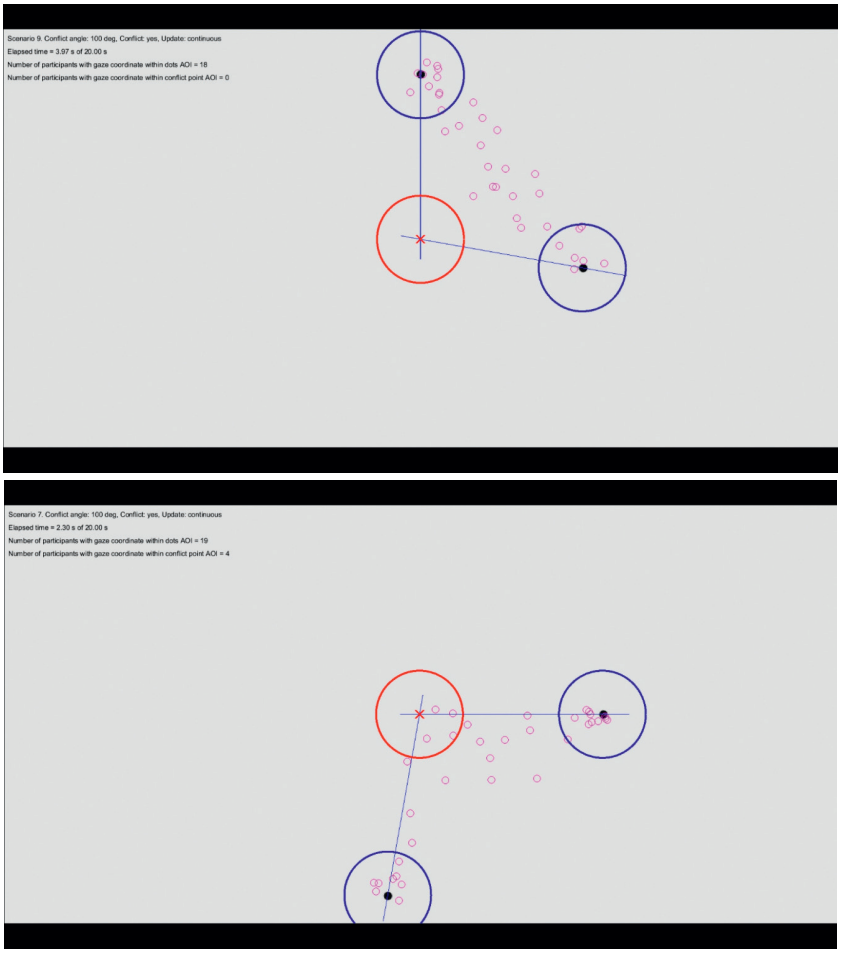


Figure 7. Snapshot from two selected scenarios (**Top**: Scenario 9, **Bottom**: Scenario 7) showing the dots (black circles), the gaze coordinates for the participants (*N* = 35), the conflict point (red X), the dots areas of interest (dots AOI, blue circles), and the conflict point area of interest (CP AOI, red circle). In the videos shown to the participants, only the two dots were visible.

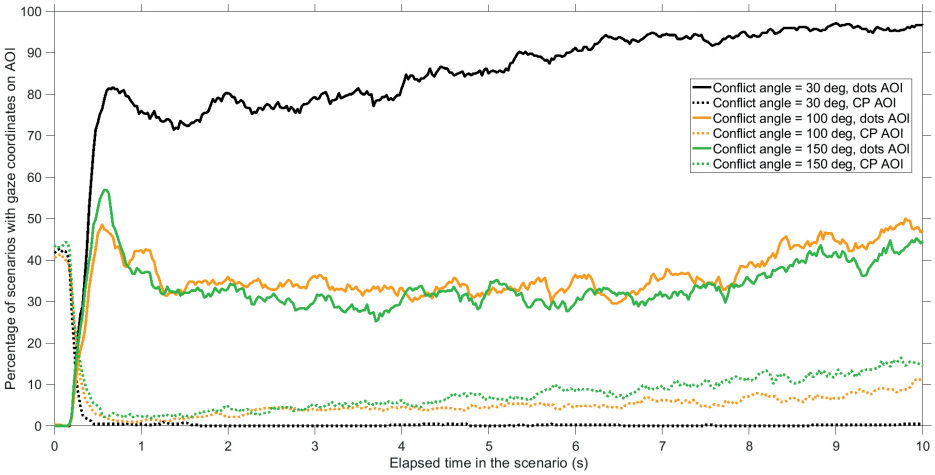


Figure 8. Percentage of participants with gaze coordinates in an area of interest (AOI) as a function of elapsed time in the scenario. A distinction is made between AOIs surrounding the dots (dots AOI) and the AOI surrounding the conflict point (CP AOI). The shown values represent averages for 35 participants and 12 scenarios per participant (conflict scenarios and non-conflict scenarios combined). For example, at an elapsed time of 4 s, for scenarios with 30 deg conflict angle, participants looked at the dots AOI in 342 of the 420 cases (81.4%), and at the CP AOI in only 1 of the 420 cases (0.2%). Only the first 10 s of the scenario are shown because from 10 s onwards, the AOIs started to overlap.

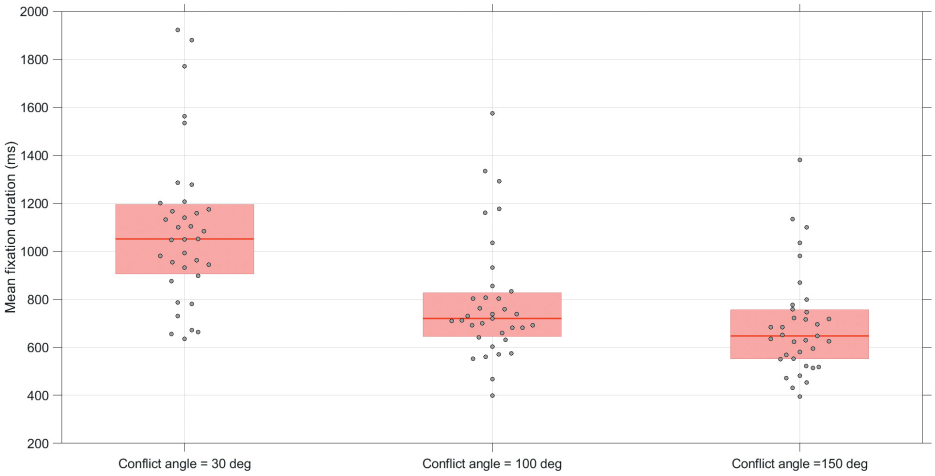


Figure 9. Boxplots of the mean fixation duration per conflict angle. The score for each participant represents the average of 12 scenarios (conflict scenarios and non-conflict scenarios combined).

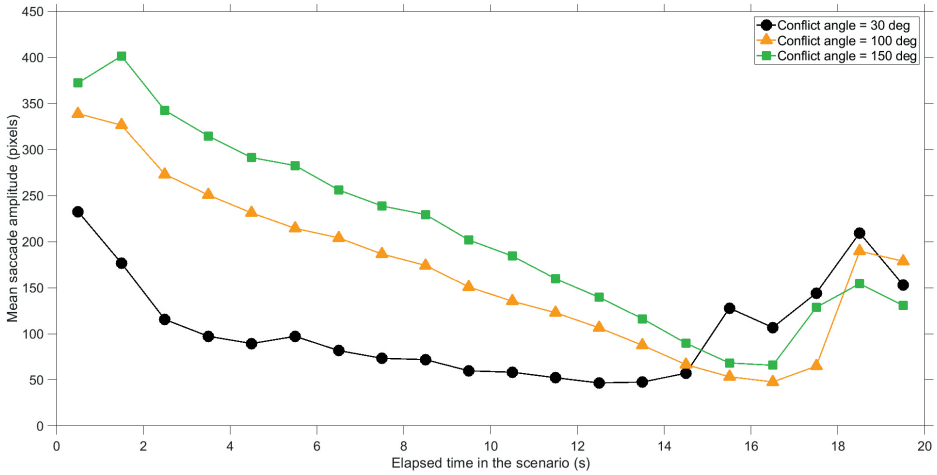


Figure 10. Mean saccade amplitude per conflict angle, where the end moments of saccades are divided into 1 s bins since the start of the scenario. The shown values represent averages for 35 participants and 12 scenarios per participant (conflict scenarios and non-conflict scenarios combined).

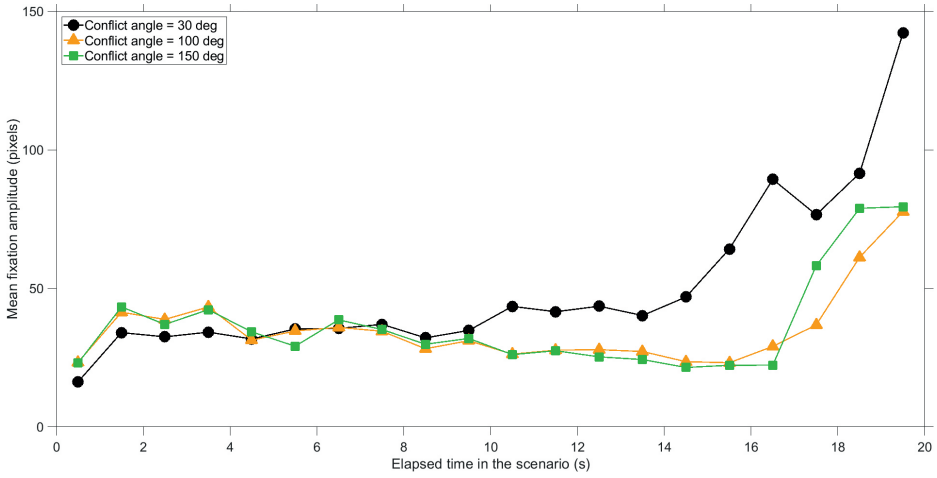


Figure 11. Mean fixation amplitude per conflict angle, where the end moments of saccades are divided into 1 s bins since the start of the scenario. The shown values represent averages for 35 participants and 12 scenarios per participant (conflict scenarios and non-conflict scenarios combined).

3.5. Scenario-Specific Effects

An overview of the dependent measures for each of the 18 scenarios is provided in the Supplementary Materials (Table S1). For most scenarios, participants distributed their attention towards Dot 1 and 2 in an approximately 50–50% manner. However, for some scenarios, participants focused more on one of the dots. In particular, if most (> 70%) of the attention went to one of the two dots, this pertained to a dot that was moving horizontally or downward.

Another noteworthy finding is that, at the level of scenarios, better performance was associated with a lower self-reported difficulty (Table S1). This relationship, which is shown in Figure 12, held for conflict scenarios (r = −0.89, n = 9) and non-conflict scenarios (r = −0.93, n = 9). In other words, participants were able to reliably assess which scenarios are more difficult than others.

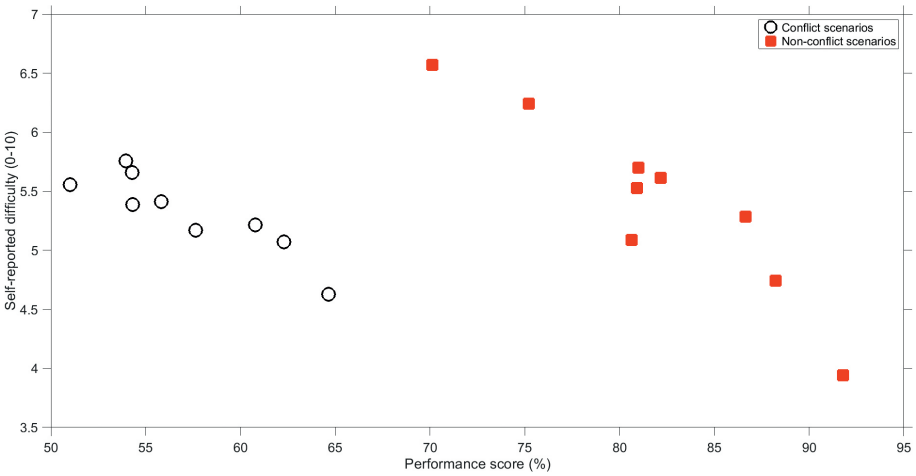


Figure 12. Mean self-reported difficulty score versus mean performance score for conflict scenarios and non-conflict scenarios. Each marker represents the average of 35 participants and 2 scenarios (discrete and continuous scenarios are combined).

4. DISCUSSION

This study aimed to investigate the effects of conflict angle on eye movements in an allocentric conflict detection task. Additionally, we studied the effects of discrete versus continuous screen-update rates on eye movements and conflict detection performance.

4.1. Effects of Conflict Angle

The results showed that conflict detection is a dynamic task in which participants' judgments become more accurate as the time to conflict decreases (Figures 3 and 4, see also Supplementary Figure S1). These findings serve as support for Neal and Kwantes [13], who argued that observers accumulate evidence over time until reaching a decision threshold.

The results showed that conflict angles of 30 deg yielded better performance and lower ratings of task difficulty than 150 deg conflict angles. In turn, 100 deg conflict angles yielded the lowest performance and were deemed the most difficult. For 30 deg angles, if there was a conflict, participants detected that conflict early, and if there was no conflict, participants were unlikely to indicate that there was. Thus, the high-performance score for the 30 deg conflict angle was because of both improved hits and reduced false positives. Figure S1 in the Supplementary Materials provides

support for these observations using an index of perceptual sensitivity (*d*'), calculated using detection theory [30]. Results for the response bias (β), also shown in the Supplementary Materials (Figure S2), indicate that participants behaved approximately as an ideal observer, that is, they assigned equal weight to Type II errors (failing to report a conflict in conflict trials) and Type I errors (reporting a conflict in non-conflict trials).

How can the superior performance for 30 deg conflict angles be explained, and why did 100 deg conflict angles yield the poorest performance? Gilden [31] argued that participants often use simple kinematic heuristics when gaining awareness of dynamical systems. The results of our study can also be explained with kinematic heuristics. For 30 deg conflict angles, it may be easy for participants to detect an imminent collision, because if one dot travels behind the other at a fixed speed, then the observer knows that the dots will not collide [11,18,21]. If the time in the scenario elapses, the relative distance of the dots to the conflict point keeps increasing, so it should become more and more evident to the observer that the trailing dot will not overtake the leading dot (see Supplementary Figure S3 for a relative distance graph). Tresilian [2] explained that this "closer is first" (p. 240) rule is easiest to apply when the two targets move in parallel. For 150 deg angles, a conflict may also be easy to detect, for example as an offset from an imaginary line connected the two converging dots [21]. For 100 deg conflict angles, however, there may have been no such kinematic rules that the participants could apply.

For 30 deg conflict angles in particular, participants employed smooth pursuit eye movements while not glancing at the future conflict point. These patterns are in agreement with the 'closer is first' strategy. By tracking the two dots, it may become apparent whether the dots move at a constant velocity and side-by-side (resulting in a collision) or that one dot lags behind the other (resulting in a safe pass). We also observed that participants preferred to look most at a dot that was moving downward or horizontally (see Section 3.5), which is consistent with literature about pursuit movements [32].

For larger conflict angles (100 and 150 deg), participants showed a higher number of fixations, and the gaze coordinates were often in between the dots and the conflict point AOIs. For these conflict angles, the dots are further apart on the screen, and observers cannot apply smooth pursuit of one dot while keeping the other dot within the foveal region. The phenomenon of looking at the conflict point can be explained by required eye-movement effort, in line with Wickens' [33] Saliency-Effort-Expectancy-Value (SEEV) model: For 150 deg conflict angles, the conflict point lies in between the two dots, making it less effort for participants to sample towards that conflict point, as compared to smaller conflict angles. In summary, the eye-movement patterns are explainable in terms of the distance between the dots, which is larger when the conflict angle is higher.

4.2. Effects of Update Rate

Continuous stimuli yielded a statistically significant improvement of conflict detection performance score as compared to discrete stimuli, with a Cohen's *d* effect size of 0.46. In other types of tasks, such as driving in a virtual driving simulator, considerably stronger effects of visual update rate have been observed. For example, Van Erp and Padmos [23] observed a factor 3 difference in lane-keeping performance between low (3 Hz) and normal (30 Hz) update rate conditions. The relatively small effects of update rate in the present study can be explained by the fact that the current task was an open-loop task in which participants did not rely on feedback to respond. The discrete presentation resulted in a delayed perception, where participants had to wait for a movement of the dot by keeping it in foveal vision in order to determine its velocity. In a closed-loop task such as car driving, the effect of a limited update rate would cause not only a delay in perception but also a delayed steering response, resulting in reduced stability of control [22].

A strong effect of update rate was found for fixation duration. That is, with discrete stimulus movement, observers fixated longer, and exhibited fewer fixations per second, as compared to continuous stimulus movement. This difference may have occurred because, with a discrete presentation of stimuli, it takes time to extract heading information. The increase of fixation duration can be interpreted as indicative of increased processing load and difficulty of interpreting the stimuli [27,34]. It is noted that in our experiment, information about the speed of the dots had to be obtained from the movement of the dots, whereas in actual applications, speed and heading information may also be available in an accompanying text label. Regardless, our results suggest that (radar) displays should update continuously rather than intermittently.

4.3. Limitations

Our task was simple, comprising of two moving dots moving at the same altitude, and a limited number of geometries of the scenarios. It is still to be determined how passing in front or behind, distance to the closest point of approach, relative speed, and the number of objects would affect attention distribution. If there are multiple moving objects, visual search for conflicts may become a crucial factor. A conflict between two moving targets may be hard to identify among multiple other moving targets, especially if the targets are far apart [15,35]. In real air traffic control, aircraft are accompanied by flight labels. The altitude labels are an important source for determining whether aircraft are in conflict [21].

Real air traffic control tasks involve multitasking, such as communication and teamwork, which in turn affect eye movements [36–38]. Furthermore, it is known that air traffic control operators tend to experience their task as safety-critical and sometimes stressful [39,40]. In our study, participants assigned about equal weight to false negatives and

false positives (see Figure S2 in the Supplementary Materials). It is expected that in real air traffic control, operators are more likely to prevent false negatives (i.e., apply a cautious strategy).

This research was conducted with engineering students. Because we tested fundamental perceptual principles, we believe that our findings are generalizable to other participant groups. However, some differences between experts and novices are to be expected. Loft et al. [3] found that air traffic control experts were more likely to intervene than trainees. Similarly, Bisseret [41] argued that experienced operators swiftly respond to a conflict, whereas trainees may feel hesitant to act once they detect a conflict. Van Meeuwen et al. [42] found that, for a task in which participants had to provide the optimal order of arrival of aircraft, expert air traffic controllers reached better solutions and applied more efficient visual scan paths as compared to novice air traffic controllers.

A final limitation is that our study was concerned with conflict detection only, with high performance in conflict trials being determined by pressing the spacebar as early as possible (Figure S4). Hilburn [43] argued that conflict resolution involves task demands that differ from conflict detection. For example, he commented that: “Similarly, head on situations seem easier to detect, but (because of high closure speed) are more difficult to resolve” (p. 57). Future research could focus on examining the interplay between conflict detection and conflict resolution.

5. CONCLUSIONS

It is concluded that conflict detection performance is better for small conflict angles (30 deg) than for near-perpendicular angles (100 deg). A small conflict angle results in pursuit movement, whereas larger conflict angles result in higher eye-movement activity and eye movements in between the dots rather than at the dots. Additionally, continuously moving stimuli yield better conflict detection performance than stimuli that moved in a discrete manner.

REFERENCES

1. Hancock, P.A.; Manser, M.P. Time-to-contact. In *Occupational Injury: Risk, Prevention and Intervention*; Feyer, A., Williamson, A., Eds.; Taylor & Francis: Bristol, CO, USA, 1998; pp. 44–58.
2. Tresilian, J.R. Perceptual and cognitive processes in time-to-contact estimation: Analysis of prediction-motion and relative judgment tasks. *Percept. Psychophys.* 1995, *57*, 231–245.
3. Loft, S.; Bolland, S.; Humphreys, M.S.; Neal, A. A theory and model of conflict detection in air traffic control: Incorporating environmental constraints. *J. Exp. Psychol. Appl.* 2009, *15*, 106–124, doi:10.1037/a0016118.
4. Matton, N.; Gotteland, J.; Granger, G.; Durand, N. Impact of ATCO training and expertise on dynamic spatial abilities. In Proceedings of the 20th International Symposium on Aviation Psychology, Dayton, OH, USA, 7–10 May 2019; pp. 385–390. Available online: https://corescholar.libraries.wright.edu/isap_2019/65 (accessed on 18 July 2020).
5. Gibson, J.J. *Motion Picture Testing and Research*; Armed Forces Aviation Psychology Program Research Reports; Report No. 7; Government Printing Office: Washington, DC, USA, 1947.
6. Rains, L.; Logg, E.; Walsh, D. *Asteroids. Computer Software*; Atari: Sunnyvale, CA, USA, 1979.
7. Foina, A.G.; Sengupta, R.; Lerchi, P.; Liu, Z.; Krainer, C. Drones in smart cities: overcoming barriers through air traffic control research. In Proceedings of the 2015 Workshop on Research, Education and Development of Unmanned Aerial Systems (RED-UAS), Cancun, Mexico, 23–25 November 2015; pp. 351–359, doi:10.1109/RED-UAS.2015.7441027.
8. d'Orey, P.M.; Hosseini, A.; Azevedo, J.; Diermeyer, F.; Ferreira, M.; Lienkamp, M. Hail-a-Drone: enabling teleoperated taxi fleets. In Proceedings of the 2016 IEEE Intelligent Vehicles Symposium (IV), Gotenburg, Sweden, 19–22 June 2016; pp. 774–781, doi:10.1109/IVS.2016.7535475.
9. Nakamura, A.; Ota, J.; Arai, T. Human-supervised multiple mobile robot system. *IEEE Trans. Robot. Autom.* 2002, *18*, 728–743, doi:10.1109/TRA.2002.803465.
10. Kimball, K.A. Estimation of intersection of two converging targets as a function of speed and angle of target movement. *Percept. Mot. Ski.* 1970, *30*, 303–310, doi:10.2466/pms.1970.30.1.303.
11. Law, D.; Pellegrino, J.; Mitchell, S.; Fischer, S.; McDonald, T.; Hunt, E. Perceptual and cognitive factors governing performance in comparative arrival-time judgements. *J. Exp. Psychol. Hum. Percept. Perform.* 1993, *19*, 1183–1199.
12. Kimball, K.A. Differential velocity and time prediction of motion. *Percept. Mot. Ski.* 1973, *36*, 935–945, doi:10.2466/pms.1973.36.3.935.
13. Neal, A.; Kwantes, P.J. An evidence accumulation model for conflict detection performance in a simulated air traffic control task. *Hum. Factors* 2009, *51*, 164–180, doi:10.1177/0018720809335071.
14. Wyndemere. *An Evaluation of Air Traffic Control Complexity*; Final report, contract number NAS 2-14284; Wyndemere: Boulder, CO, USA, 1996.
15. Mackintosh, M.-A.; Dunbar, M.; Lozito, S.; Cashion, P.; Mcgann, A.; Dulchinos, V.; Van Gent, R. Self-separation from the air and ground perspective. In Proceedings of the 2nd USA/Europe Air Traffic Management R&D Seminar, Orlando, FL, USA, 1–4 December 1998.

16. Pompanon, C.; Raufaste, É. Extrapolation of the intersection of two trajectories on a 2D display: Evidence of biases. In Proceedings of the 2007 International Symposium on Aviation Psychology, Dayton, OH, USA, 23–26 April 2007; pp. 530–536.

17. Marchitto, M.; Benedetto, S.; Baccino, T.; Cañas, J.J. Air traffic control: Ocular metrics reflect cognitive complexity. *Int. J. Ind. Ergon.* 2016, *54*, 120–130, doi:10.1016/j.ergon.2016.05.010.

18. Xu, X.; Rantanen, E.A. Conflict detection in air traffic control: A task analysis, a literature review, and a need for further research. In Proceedings of the 12th International Symposium on Aviation Psychology, Dayton, OH, USA, 14–17 April 2003; pp. 1289–1295.

19. Just, M.A.; Carpenter, P.A. A theory of reading: From eye fixations to comprehension. *Psychol. Rev.* 1980, *87*, 329–354, doi:10.1037/0033-295X.87.4.329.

20. Hunter, A.C.; Parush, A. Using eye movements to uncover conflict detection strategies. *Proc. Hum. Factors Ergon. Soc.* 2009, *3*, 1729–1733, doi:10.1518/107118109x12524444081872.

21. Pompanon, C.; Raufaste, É. The intervention trigger model: Computational modelling of air traffic control. *Proc. Annu. Meet. Cogn. Sci. Soc.* 2009, *31*, 2262–2267.

22. Chen, J.Y.C.; Thropp, J.E. Review of low frame rate effects on human performance. *IEEE Trans. Syst. Man Cybern. Part A Syst. Hum.* 2007, *37*, 1063–1076, doi:10.1109/TSMCA.2007.904779.

23. Van Erp, J.B.F.; Padmos, P. Image parameters for driving with indirect viewing systems. *Ergonomics* 2003, *46*, 1471–1499, doi:10.1080/0014013032000121624.

24. Thomas, L.C.; Wickens, C.D. Display dimensionality, conflict geometry, and time pressure effects on conflict detection and resolution performance using cockpit displays of traffic information. *Int. J. Aviat. Psychol.* 2006, *16*, 321–342, doi:10.1207/s15327108ijap1603_5.

25. Rantanen, E.M.; Nunes, A. Hierarchical conflict detection in air traffic control. *Int. J. Aviat. Psychol.* 2005, *15*, 339–362, doi:10.1207/s15327108ijap1504.

26. Eisma, Y.B.; Cabrall, C.D.D.; De Winter, J.C.F. Visual sampling processes revisited: Replicating and extending Senders (1983) using modern eye-tracking equipment. *IEEE Trans. Hum. Mach. Syst.* 2018, *48*, 526–540, doi:10.1109/THMS.2018.2806200.

27. Underwood, G.; Crundall, D.; Chapman, P. Driving simulator validation with hazard perception. *Transp. Res. Part F Traffic Psychol. Behav.* 2011, *14*, 435–446, doi:10.1016/j.trf.2011.04.008.

28. Holmqvist, K.; Nyström, M.; Andersson, R.; Dewhurst, R.; Jarodzka, H.; Van de Weijer, J. *Eye tracking: A Comprehensive Guide to Methods and Measures*; OUP: Oxford, UK, 2011.

29. Faul, F.; Erdfelder, E.; Lang, A.G.; Buchner, A. G * Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* 2007, *39*, 175–191, doi:10.3758/BF03193146.

30. Abdi, H. Signal detection theory (SDT). In *Encyclopedia of Measurement and Statistics*; Sage: Thousands Oaks, CA, USA, 2007; pp. 886–889. Available online: https://wwwpub.utdallas.edu/~herve/abdi-SDT_2009.pdf (accessed on 18 July 2020).

31. Gilden, D.L. On the origins of dynamical awareness. *Psychol. Rev.* 1991, *98*, 554–568, doi:10.1037//0033-295x.98.4.554.

32. Ke, S.R.; Lam, J.; Pai, D.K.; Spering, M. Directional asymmetries in human smooth pursuit eye movements. *Investig. Ophthalmol. Vis. Sci.* 2013, *54*, 4409–4421, doi:10.1167/iovs.12-11369.

33. Wickens, C.D. Visual attention control, scanning, and information sampling. In *Applied Attention Theory*; Wickens, C.D., McCarley, J.S., Eds.; CRC Press: Boca Raton, FL, USA, 2008; pp. 41–61, doi:10.1201/9781420063363.ch4.

34. Fitts, P.M.; Jones, R.E.; Milton, J.L. Eye movements of aircraft pilots during instrument-landing approaches. *Aeronaut. Eng. Rev.* 1950, *9*, 1–6.

35. Remington, R.W.; Johnston, J.C.; Ruthruff, E.; Gold, M.; Romera, M. Visual search in complex displays: Factors affecting conflict detection by air traffic controllers. *Hum. Factors* 2000, *42*, 349–366, doi:10.1518/001872000779698105.

36. Li, W.C.; Kearney, P.; Braithwaite, G.; Lin, J.J.H. How much is too much on monitoring tasks? Visual scan patterns of single air traffic controller performing multiple remote tower operations. *Int. J. Ind. Ergon.* 2018, *67*, 135–144, doi:10.1016/j.ergon.2018.05.005.

37. Metzger, U.; Parasuraman, R. Effects of automated conflict cuing and traffic density on air traffic controller performance and visual attention in a datalink environment. *Int. J. Aviat. Psychol.* 2006, *16*, 343–362, doi:10.1207/s15327108ijap1604_1.

38. Willems, B.; Allen, R.C.; Stein, E.S. *Air Traffic Control Specialist Visual Scanning II: Task Load, Visual Noise, and Intrusions Into Controlled Airspace*; Technical Report; Federal Aviation Administration, William J. Hughes Technical Center, Washington, DC, USA, 1999. Available online: https://apps.dtic.mil/dtic/tr/fulltext/u2/a372988.pdf (accessed on 18 July 2020).

39. Finkelman, J.M.; Kirschner, C. An information-processing interpretation of air traffic control stress. *Hum. Factors* 1980, *22*, 561–567, doi:10.1177/001872088002200505.

40. Zeier, H.; Brauchli, P.; Joller-Jemelka, H.I. Effects of work demands on immunoglobulin A and cortisol in air traffic controllers. *Biol. Psychol.* 1996, *42*, 413–423, doi:10.1016/0301-0511(95)05170-8.

41. Bisseret, A. Application of signal detection theory to decision making in supervisory control: The effect of the operator's experience. *Ergonomics* 1981, *24*, 81–84, doi:10.1080/00140138108924833.

42. Van Meeuwen, L.W.; Jarodzka, H.; Brand-Gruwel, S.; Kirschner, P.A.; De Bock, J.J.P.R.; Van Merriënboer, J.J.G. Identification of effective visual problem solving strategies in a complex visual domain. *Learn. Instr.* 2014, *32*, 10–21, doi: 10.1016/j.learninstruc.2014.01.004.

43. Hilburn, B. *Cognitive Complexity in Air Traffic Control: A Literature Review*; Ph.D. Thesis, Center for Human Performance Research, Sacramento, CA, USA, 2004. Available online: https://www.eurocontrol.int/cognitive-complexity-air-traffic-control-literature-review (accessed on 18 July 2020).

Table S1. Means (color-coded) and standard deviations of dependent variables per scenario (*N* = 35).

Scenario number	Conflict angle (deg)	Conflict	1. Fixation rate (Hz)		2. Mean fixation duration (ms)		3. Mean saccade amplitude (pixels)		3. Mean fixation amplitude (pixels)		Proportion of time on Dot 1		Proportion of time on dots AOI		Proportion of time on CP AOI		Performance score (%)		Self-reported difficulty (0-10)		Self-reported difficulty (0-10)		Number of spacebar presses (#)	
1	30	Yes	0.566	0.218	1194	452	126	38	39	15	0.750	0.139	0.951	0.039	0.262	0.050	60.772	15.909	5.214	1.655	5.214	1.655	1.01	0.08
2	30	Yes	0.674	0.289	1000	440	106	26	35	15	0.558	0.175	0.947	0.031	0.262	0.036	64.653	13.664	4.629	1.601	4.629	1.601	1.01	0.08
3	30	Yes	0.726	0.264	1041	479	112	30	38	17	0.732	0.132	0.932	0.063	0.275	0.042	62.305	15.904	5.071	1.520	5.071	1.520	1.01	0.08
4	30	No	0.670	0.301	1245	523	121	41	42	19	0.232	0.138	0.952	0.048	0.237	0.049	80.625	15.856	5.086	1.881	5.086	1.881	0.60	0.42
5	30	No	0.711	0.246	1111	473	155	65	39	14	0.503	0.160	0.946	0.025	0.250	0.035	86.654	14.155	5.286	2.136	5.286	2.136	0.44	0.40
6	30	No	0.896	0.365	977	359	110	23	40	17	0.636	0.163	0.937	0.052	0.283	0.046	91.792	11.124	3.943	1.748	3.943	1.748	0.26	0.33
7	100	Yes	1.116	0.306	783	259	187	44	32	15	0.707	0.090	0.711	0.123	0.399	0.094	50.999	19.775	5.557	2.068	5.557	2.068	1.11	0.27
8	100	Yes	1.226	0.365	723	268	215	45	38	20	0.625	0.111	0.739	0.131	0.335	0.048	55.816	18.447	5.414	1.873	5.414	1.873	1.07	0.18
9	100	Yes	1.144	0.388	788	365	191	52	36	17	0.529	0.128	0.671	0.153	0.345	0.052	53.956	17.928	5.757	1.729	5.757	1.729	1.03	0.12
10	100	No	1.211	0.330	752	215	184	46	28	13	0.655	0.133	0.687	0.124	0.401	0.093	70.112	14.323	6.571	1.582	6.571	1.582	0.86	0.38
11	100	No	1.121	0.373	878	395	176	42	36	15	0.742	0.117	0.732	0.131	0.318	0.056	82.180	16.101	5.614	1.922	5.614	1.922	0.57	0.39
12	100	No	1.183	0.379	806	270	187	48	37	17	0.604	0.129	0.680	0.150	0.314	0.059	75.206	15.799	6.243	1.804	6.243	1.804	0.76	0.46
13	150	Yes	1.324	0.305	648	178	243	59	30	17	0.555	0.125	0.659	0.147	0.458	0.114	54.295	14.676	5.657	1.748	5.657	1.748	1.09	0.23
14	150	Yes	1.378	0.348	638	203	228	49	34	22	0.665	0.099	0.685	0.151	0.445	0.096	57.644	17.003	5.171	1.761	5.171	1.761	1.04	0.19
15	150	Yes	1.200	0.381	735	304	219	53	39	20	0.629	0.107	0.645	0.148	0.459	0.090	54.298	16.902	5.386	1.871	5.386	1.871	1.07	0.18
16	150	No	1.334	0.366	735	278	237	50	35	21	0.582	0.121	0.670	0.145	0.478	0.133	88.244	12.186	4.743	1.550	4.743	1.550	0.47	0.42
17	150	No	1.397	0.426	690	307	233	50	32	17	0.636	0.120	0.705	0.142	0.451	0.107	80.997	16.776	5.700	1.733	5.700	1.733	0.56	0.43
18	150	No	1.371	0.390	711	320	218	52	35	20	0.669	0.092	0.634	0.147	0.442	0.088	80.906	17.119	5.529	1.863	5.529	1.863	0.61	0.46

Note. The results for each participant were averaged for the 18 continuous scenarios and the 18 discrete scenarios. The additional measure 'proportion of time on Dot 1' represents the proportion of time that the gaze coordinate was closer to Dot 1 than to Dot 2. CP = conflict point.

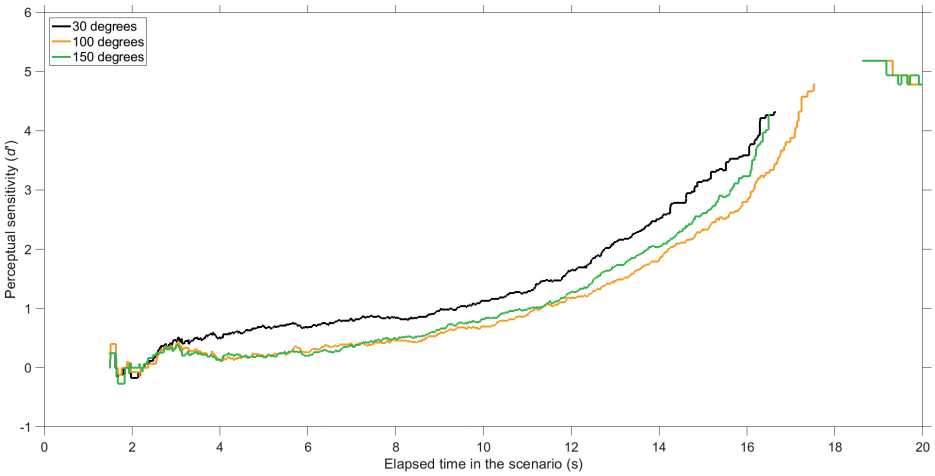


Figure S1. Perceptual sensitivity (*d'*) as a function of elapsed time during the scenario, calculated from the results shown in Figures 3 and 4. It can be seen that perceptual sensitivity is highest for 30 deg conflict angles. Also, perceptual sensitivity increases with elapsed time, which can be explained because it gradually becomes evident whether or not a collision will occur.

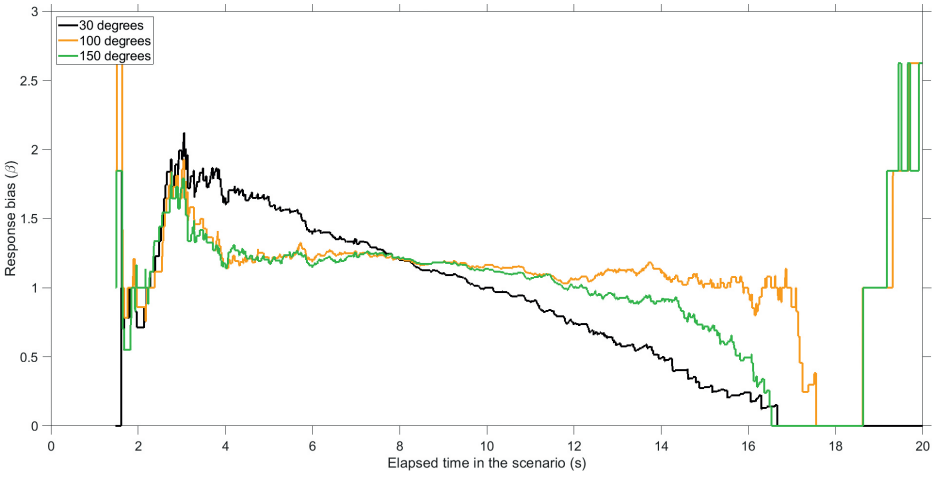


Figure S2. Response bias (β) as a function of elapsed time, calculated from the results shown in Figure 3 (showing hit rates) and Figure 4 (showing false alarm rates). β = 1 would represent an 'ideal observation' where the miss rate equals the false positive rate. It can be seen that β is about 1 for 100 deg and 150 deg conflict angles, whereas β decreases with elapsed time for 30 deg conflict angles. To illustrate, at about 16 seconds into the 30-deg scenarios, the miss rate was low (1%, or 99% hit rate) but the false alarm rate was high (8%), indicating that participants were cautious (i.e., liberal, low β) at that point in time. In other words, in non-conflict scenarios, some participants kept pressing the spacebar to indicate that the dots could collide even when the dots would not collide, an effect that may be due to a delay in the human response.

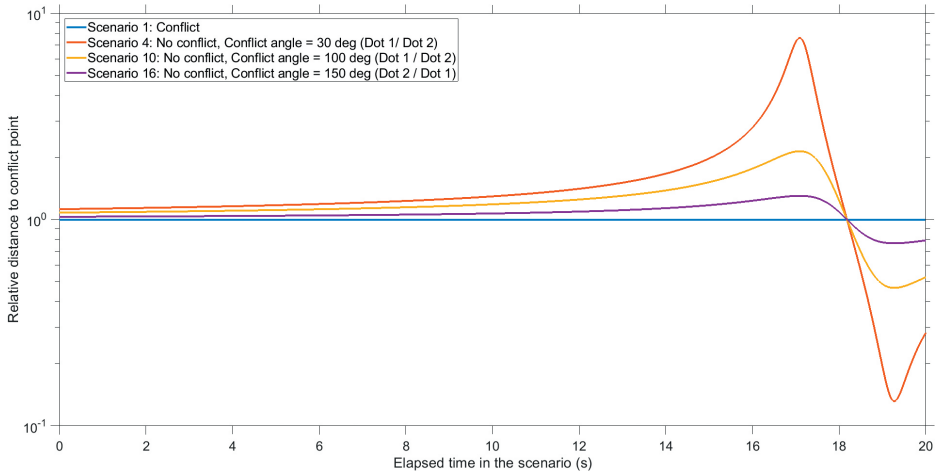


Figure S3. Ratio between the distance from Dot 1 to the conflict point and the distance from Dot 2 to the conflict point. For example, if the value equals 2, then one dot is twice as far from the conflict point as the other dot. Note that the closest point of approach is at 18.3 s.

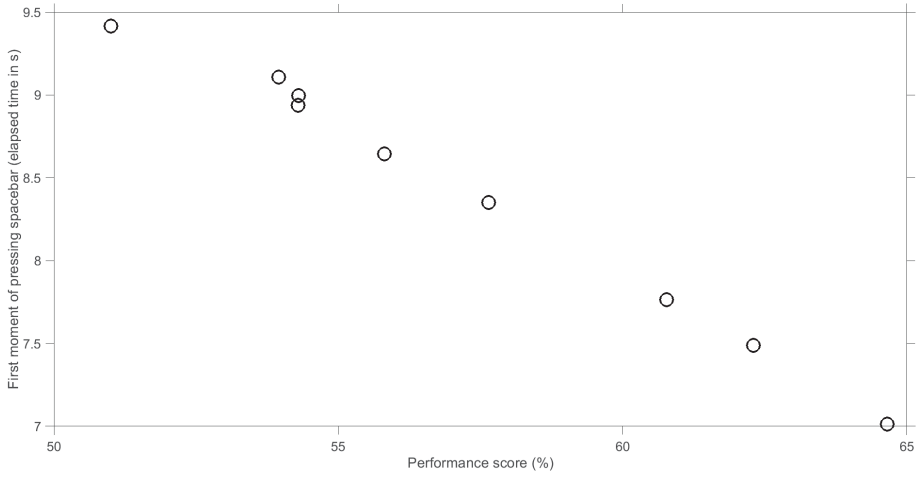


Figure S4. Mean first moment of pressing the spacebar versus mean performance score for conflict scenarios. Each marker represents the average of 35 participants and 2 scenarios (discrete and continuous scenarios are combined). The strong correlation indicates that the moment of pressing the spacebar and the performance score are redundant variables at the level of scenarios.

CHAPTER 6

Augmented Visual Feedback: Cure or Distraction?

Eisma, Y. B., Borst, C., Van Paassen, M. M., & De Winter, J. C. F. (in press). Augmented visual feedback: Cure or distraction? *Human Factors.*

ABSTRACT

Objective. To investigate the effect of augmented feedback on participants' workload, performance, and distribution of visual attention.

Background. An important question in human-machine interface design is whether the operator should be provided with direct solutions. We focused on the solution space diagram (SSD), a type of augmented feedback that shows directly whether two aircraft are on conflicting trajectories.

Methods. One group of novices (*n* = 13) completed conflict detection tasks with SSD, whereas a second group (*n* = 11) performed the same tasks without SSD. Eye-tracking was used to measure visual attention distribution.

Results. The mean self-reported task difficulty was substantially lower for the SSD group compared to the No-SSD group. The SSD group had a better conflict detection rate than the No-SSD group, whereas false-positive rates were equivalent. High false-positive rates for some scenarios were attributed to participants who misunderstood the SSD. Compared to the No-SSD group, the SSD group spent a large proportion of their time looking at the SSD aircraft while looking less at other areas of interest.

Conclusions. Augmented feedback makes the task subjectively easier, but has side effects related to visual tunneling and misunderstanding.

Application. Caution should be exercised when human operators are expected to reproduce task solutions that are provided by augmented visual feedback.

INTRODUCTION

Automation is present in many aspects of society, including areas such as process control, human transportation (e.g., driverless metro trains), and warehouse logistics. However, in complex work domains such as air traffic control, anesthesia care, and car driving, full automation is not yet feasible because of the high risks involved (Bazilinskyy, Kyriakidis, Dodou, & De Winter, 2019; Kaber & Endsley, 2004; Parasuraman, Sheridan, & Wickens, 2000). Although information acquisition and analysis are highly automated, final decision-making is left to a human operator. In air traffic control, for example, a human controller supervises radar screens to decide which routing instructions to give to pilots in order to structure the airflow safely and efficiently (Sheridan, 2002).

A crucial question for the above domains is what information should be shown on the display, and what visual appearance the information should have. One approach would be to present all the data that the operator might need. However, as explained by Sheridan (1995), "humans can absorb and make use of only very limited quantities of information. It is well established that displaying all the information that might be useful means there is too much information to be able to find what is needed." (p. 825). Another approach, which is the focus of the current paper, would be to let the computer transform the available sensor data into intuitive visualizations for the task at hand. This approach may be attractive for systems designers who may want to ensure maximal operator compliance. However, this approach may involve risks in the unlikely case that the provided solution is invalid, for example, in cases where vital sensor data is missing or incorrect. Thus, a potential disadvantage of providing operators with augmented feedback or other types of guidance is that operators 'blindly' follow the suggested action without checking task-relevant elements of the work domain (Parasuraman, Molloy, & Singh, 1993). As pointed out by Sheridan (2002), the use of a decision aid implies that the "human can properly decide when the situation includes elements the decision aid can properly assess and can know for which elements the decision aid should be ignored." (p. 150).

The hypothesized risk of decision aids corresponds to theories about 'guidance effects' of augmented feedback as studied in the area of motor learning. Wulf and Shea (2004), for example, stated that concurrent augmented feedback "typically has very strong performance-enhancing effects" (p. 128). However, they also noted that, compared to post-trial feedback, concurrent feedback is expected to result in a performance decrement when the feedback is removed. Schmidt and Wulf (1997) argued that concurrent feedback distracts attention from task-intrinsic feedback. Here, intrinsic task feedback is defined as the natural cues in the work environment that are necessary for executing the task correctly, in the absence of augmented feedback.

In the present study, we employed a display called the Solution Space Diagram (SSD) (Bijsterbosch, Borst, Mulder, & Van Paassen, 2016). The SSD, which has been used in air traffic control (ATC) research, shows the operator whether the current situation is safe or unsafe based on whether the aircraft's speed vector resides in a no-go zone (a red triangle). In case of a conflict between two aircraft, the operator can reposition the speed vector outside of the no-go zone to resolve a conflict. It is known that ATC operators normally tend to resolve conflicts between aircraft through heading control, whereas speed control seems an underused strategy (Ehrmanntraut, 2004; Hilburn, Westin, & Borst, 2014). The SSD shows the operator the entire solution space, and therefore facilitates speed control as well as heading control.

Previous research showed that the SSD contributes to reduced self-reported workload during an ATC task as compared to no SSD (Mercado-Velasco, Mulder, & Van Paassen, 2010). However, it is unknown whether participants who use the SSD may be distracted from processing task-intrinsic cues such as the state of other aircraft shown on the screen. Herein, we used eye-tracking to test Schmidt and Wulf's (1997) hypothesis that augmented feedback guides attention away from task-intrinsic cues. Thus, besides verifying whether the SSD results in performance improvements (fewer misses and false alarms) and lower self-reported workload as compared to not using the SSD, we examined how participants distributed their visual attention across the display.

METHODS

Participants

The participants were twenty-four engineering MSc and PhD students. Their mean age was 24.6 years (*SD* = 4.3 years). The SSD group consisted of 12 males and 1 female and had a mean age of 24.2 years (*SD* = 3.2). The No-SSD group consisted of 10 males and 1 female and had a mean age of 25.0 years (*SD* = 5.2). Participants were allocated in a random manner between the two groups. Ten participants were recruited from the faculty of Aerospace Engineering; the remaining 14 participants were recruited from the faculty of Mechanical Engineering. For the Aerospace Engineering participants, we asked whether the participant was already familiar with the SSD (e.g., from a lecture or research). Two participants who indicated being familiar with the SSD were allocated to the No-SSD group.

This research complied with the American Psychological Association Code of Ethics and was approved by the Human Research Ethics Committee at the Delft University of Technology. Informed consent was obtained from each participant.

Procedures and Task

First, participants provided their age and gender. Next, they received general instructions, stating: "In this experiment you are asked to perform a *conflict detection*

task. You are presented with static Air Traffic Control (ATC) scenarios, each containing two aircraft. For each scenario we need your judgment of whether the two aircraft are on conflicting trajectories, or not. In case the aircraft are in conflict, the aircraft will collide in the future. In case the aircraft are not in conflict, the aircraft will pass by. It is your task to *press the spacebar* if you think the two aircraft are in conflict. In case you think that the aircraft are not in conflict, then do nothing. You are presented with 44 ATC scenarios. Each scenario will last 10 seconds."

Participants from the No-SSD group and the SSD group were shown a conflict scenario without SSD, and the following text: "Here, you see two aircraft represented by square markers. The tip of the black line in front of the marker indicates the future position of the aircraft after one minute. This scenario **does contain a conflict**. It is your job to **press the spacebar** when you think the aircraft are in conflict. If you think there is no conflict, **then do nothing**."

This screen was then followed by a screen containing a non-conflict scenario, and the following text: "Here another example is given. This scenario does **not** contain a conflict."

Participants from the SSD group received two extra instruction screens with information about how the SSD worked. First, they were shown the same conflict scenario as before, but now with SSD. The accompanying text said: "In 36 of the trials you are supported by the Solution Space Diagram (SSD). The SSD consists of two circles: The small circle represents the minimum speed of the aircraft (the shortest the speed vector can get); the larger circle indicates the maximum speed of the aircraft (the longest the speed vector can get). The red shape indicates the no-go zone, related to the intruder aircraft. If the tip of the speed vector points into the red triangle, both aircraft are in conflict. **This scenario does contain a conflict.** It is your job to **press the spacebar** when you think the aircraft are in conflict. If you think there is no conflict, **then do nothing**."

On the next screen, participants from the SSD group were shown the same non-conflict scenario as before, now with SSD support. The accompanying text said: "Here another example is given. This scenario does **not** contain a conflict."

Next, a calibration of the eye tracker was performed, after which the experiment started. The participants then viewed 44 scenarios, each for 10 seconds. Participants were presented with 36 regular scenarios (3 conflict angles x 2 conflict outcomes, each combination in 6 different configurations) and 8 transfer scenarios (4 conflict angles x 2 conflict outcomes). The transfer scenarios featured no SSD and conflict angles that were different from the conflict angles in the regular scenarios (see Section Design of the Stimuli). Table I provides an overview of the design of the experiment. The order in which the scenarios were presented was identical for every participant.

The transfer scenarios were included as an extra feature, with the aim to measure short-term transfer of learning. Because of our limited sample size and limited statistical power, we refrained from a detailed analysis of the transfer trials. Results in this paper are all based on the regular trials; the results regarding the transfer trials can be found in the Supplementary Materials. The transfer results may be useful for defining and designing future research on this topic.

Table I. Overview of the scenarios for the two experimental groups

	No-SSD group	**SSD group**
Regular scenarios: 1–18 (11 conflicts, 7 non-conflicts)	No SSD	SSD
Transfer scenarios: 19–22 (4 conflicts, 0 non-conflicts)	No SSD	No SSD
Regular scenarios: 23–40 (7 conflicts, 11 non-conflicts)	No SSD	SSD
Transfer scenarios: 41–44 (0 conflicts, 4 non-conflicts)	No SSD	No SSD

The scenarios all displayed two aircraft on converging tracks. After each scenario, participants rated the difficulty of the preceding trial, by answering the statement "The task was difficult" on a scale of 0 (completely disagree) to 10 (completely agree). The experiment lasted about 15 minutes per person.

Apparatus

Eye movements were recorded at 2000 Hz using the SR-Research Eyelink 1000 Plus. The eye-tracker featured binocular measurements. However, binocular tracking was not always available due to the loss of tracking of one eye. The recorded gaze coordinates of the left and right eye were averaged if left and right were both available.

The stimuli were displayed on a 24-inch BENQ monitor with a resolution of 1920 x 1080 pixels (531 x 298 mm). The refresh rate of the monitor was 60 Hz. The distance between the monitor and the head support was approximately 95 cm, and the distance between the eye-tracking camera/IR light source was approximately 65 cm. The monitor suspended a horizontal and vertical viewing angle of 31 deg and 18 deg, respectively.

Independent Variables

The first independent variable was the availability of the SSD. This was a between-subjects variable. The second independent variable was the conflict outcome. In half of the scenarios, there was a conflict, and in the other half, there was no conflict. In non-conflict scenarios, the distance between aircraft during the closest point of approach (CPA) was 7 Nautical Miles (abbreviated: NM; 112 pixels or 1.87 deg on the screen); in conflict scenarios, the closest point of approach was 0 NM. The conflict outcome was a within-subject variable.

Design of the Stimuli

The scenarios were static ATC images with a resolution of 1920 x 1080 pixels. Each scenario featured two aircraft. An aircraft was represented by a square marker with a speed vector (black line) indicating the predicted traveled distance over 1 minute, which at a speed of 245 knots corresponds to 4.1 NM or 65 pixels (1.08 deg) on the screen. Thus, a distance of 1 NM corresponded to 16 pixels (0.27 deg) on the screen. Figure 1 shows one scenario without and with SSD.

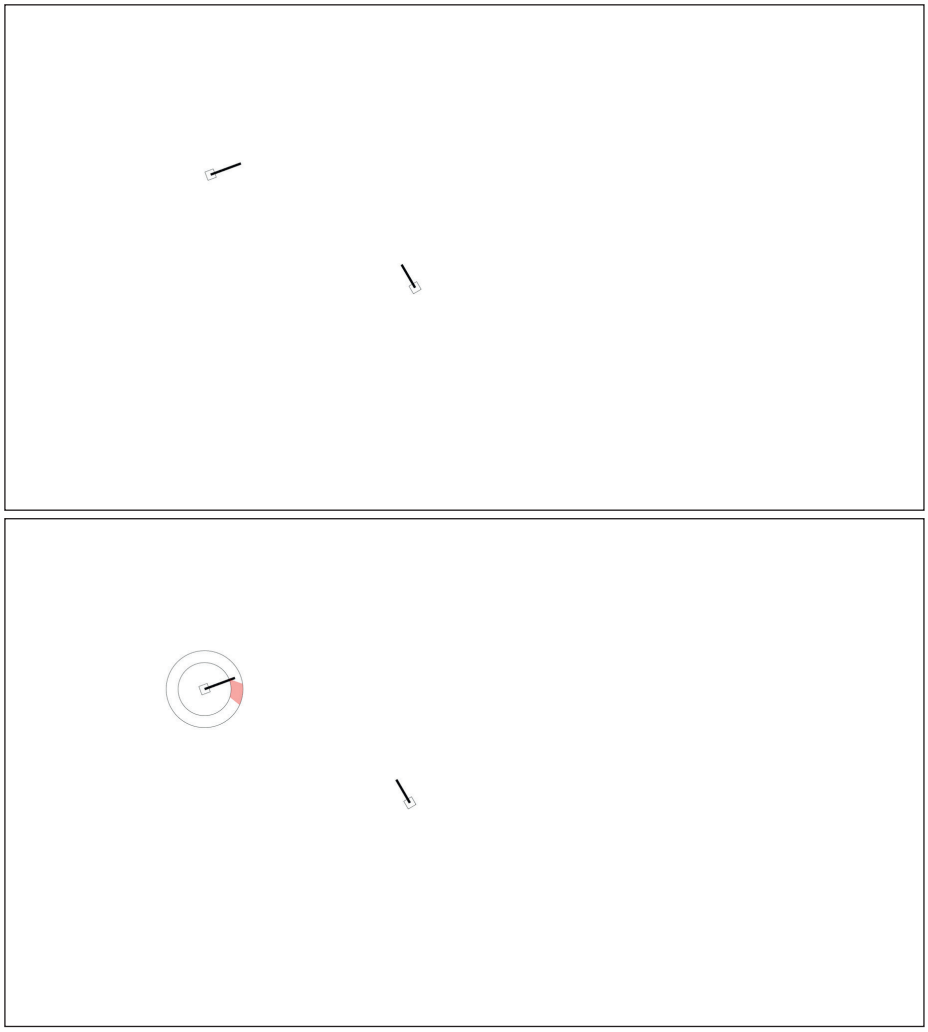


Figure 1. One of the scenarios without conflict (Scenario #10). The conflict angle is 100 deg. Top: No SSD, Bottom: SSD. If the tip of the speed vector resides in the red zone, the two aircraft are in conflict. The two concentric circles indicate the minimum and maximum speed of the aircraft.

In 22 of the scenarios, the aircraft were in conflict, which meant that a loss of separation would occur after 5 minutes and that the aircraft would collide. A loss of separation was defined as the moment the distance between the two aircraft dropped below 5 NM (80 pixels, 1.33 deg). In the other 22 scenarios, the aircraft were not in conflict, which meant that the aircraft safely passed by after 5 minutes. The closest distance for non-conflict aircraft scenarios was 7 NM (112 pixels, 1.87 deg). This closest distance of 7 NM was based on pilot tests, where we aimed for an intermediate level of difficulty. That is, we wanted participants to score better than chance (higher than 50% correct performance) but not obtain perfect performance (i.e., lower than 100% correct performance).

Thomas and Wickens (2006) defined three categories of conflict angle between aircraft: (1) overtake: 0–60 deg, (2) crossing: 60–120 deg, and (3) head-on: 120–180 deg. For this experiment, one conflict angle from each of these categories was used. Specifically, we used 30, 100, and 150 deg (12 scenarios per conflict angle). The transfer scenarios had conflict angles of 15, 35, 65, and 145 deg (2 scenarios per conflict angle).

The task was two-dimensional, with the two aircraft flying at the same altitude. The speed of Aircraft 1 (i.e., the aircraft which could potentially contain the SSD) was 245 knots, whereas the speed of Aircraft 2 ranged between 200 and 290 knots. This speed variation between scenarios was implemented to ensure that the scenarios were not perceived as simple geometrical problems. The heading and position of Aircraft 1 (and therefore Aircraft 2) was different for each scenario and obtained using a random number generator. All participants viewed the same 44 scenarios in the same order.

Dependent Variables

A non-causal median filter with a 100 ms interval was used to cancel out high-frequency camera noise while preserving the information embedded in rapid saccades (see also Eisma, Cabrall, & De Winter, 2018). Fixations and saccades were extracted using a standard filter (Eisma et al., 2018). Missing data due to blinks were linearly interpolated. The dependent variables were defined as follows:

- *Self-reported difficulty (0–10).* A difficulty score between 0 (completely disagree) and 10 (completely agree) was provided by the participants after each scenario.
- *Correct detection (%).* The percentage of conflict scenarios for which the participant pressed the spacebar.
- *Correct detection response time (RT) (ms).* The mean spacebar response time for conflict scenarios.
- *False positives (%).* The percentage of non-conflict scenarios for which the participant pressed the spacebar.

- *Mean fixation duration (s).* During fixations, participants acquire information from the visual array. For calculating the fixation duration, the eye-tracking data were partitioned into saccades and fixations, as in Eisma et al. (2018). First, the gaze speed was filtered with a Savitzky-Golay filter with order 2 and a frame length of 41. A saccade velocity threshold of 2000 pixels per second was used. The minimum fixation duration was set at 40 ms.
- *Mean saccade amplitude (pixels).* Saccade amplitude is another common measure in eye-tracking research (Underwood, Crundall, & Chapman, 2011). A higher mean saccade amplitude indicates that participants have a broader spread of fixations on the screen.
- *Gaze coordinates on area of interest (AOI) (% of time).* We computed the percentage of the total fixation time the participants fixated on (1) Aircraft 1 (possibly containing the SSD), (2) Aircraft 2 (never containing an SSD), (3) the conflict point (CP), or (4) along the lines connecting the aircraft and the CP. For Aircraft 1, Aircraft 2, and the CP, a circle of 100-pixel radius (1.67 deg) was used as a boundary of the AOI. For the connecting lines, a maximum distance to the lines of 50 pixels (0.83 deg) was used to bound the AOI. The sizes of these AOIs were based on a prior conflict detection task using the same eye tracker (Eisma, Looijestijn, & De Winter, 2019). The use of circles of 100-pixel radius ensured sufficient separation of AOIs (i.e., no misclassifications due to no overlap).

Differences between the SSD and No-SSD group were compared using independent-samples *t*-tests. An alpha value of 0.05 was used. The reason for using *t*-tests as opposed to multivariate tests was that we wanted to assess the effect of each dependent variable separately.

RESULTS

The results in this section are for the regular scenarios (Scenarios 1–18, 23–40). The results for the transfer scenarios can be found in the Supplementary Materials. Table 2 shows that participants from the SSD group found the task considerably easier than participants from the No-SSD group. These results are illustrated using Figure 2.

Table II. Means (standard deviations in parentheses) of dependent variables for the No-SSD group and the SSD group during the regular scenarios. Also shown are the results for independent-samples t-tests.

	Regular scenarios			
	No SSD (*n* = 13)	**SSD (*n* = 11)**	***t***	***p***
Difficulty (0 to 10)	4.56 (0.94)	1.53 (1.41)	6.26	**< .001**
Correct detection (%)	79.1 (11.1)	93.4 (15.3)	-2.66	**.014**
Correct detection RT (ms)	4577 (1285)	2535 (1384)	3.75	**.001**
False positive (%)	17.5 (9.8)	14.6 (17.4)	0.51	.617
Saccade amplitude (px)	216 (31)	214 (22)	0.19	.850
Fixation duration (ms)	525 (61)	794 (199)	-4.64	**< .001**
Fixations Aircraft 1 (% of time)	29.4 (5.6)	57.1 (10.4)	-8.27	**< .001**
Fixations Aircraft 2 (% of time)	25.3 (5.9)	13.2 (3.9)	5.83	**< .001**
Fixations CP (% of time)	8.9 (3.9)	5.2 (5.6)	1.90	.070
Fixations lines (% of time)	17.3 (5.1)	9.5 (3.7)	4.20	**< .001**

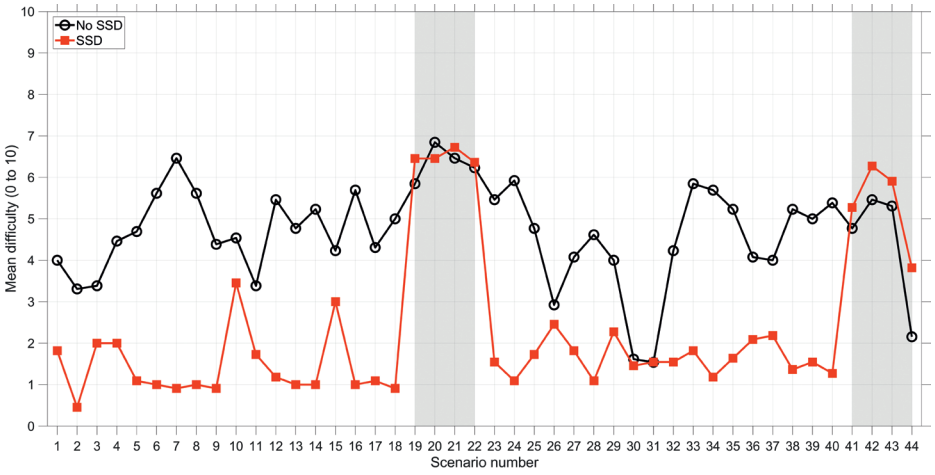


Figure 2. Mean self-reported difficulty as a function of scenario number. Scenarios 19–22 and 41–44 are transfer scenarios.

Participants from the SSD group showed a higher conflict detection rate (i.e., more often pressed the spacebar) than participants from the No-SSD group, a statistically significant difference. Participants from the SSD group also detected conflicts significantly faster than the No-SSD participants (Table II). For non-conflict scenarios, there was no significant difference between the SSD group and the No-SSD group. In other words, the SSD increased correct detections but did not diminish false positives.

As mentioned above, the SSD did not yield a significantly diminished false positive rate compared to the No-SSD group, even though the SSD always correctly indicated that the scenario was a no-conflict scenario. To better understand this finding, we explored for which type of scenarios, participants had a high false-positive rate while using the SSD. From the 18 non-conflict scenarios, six were of a special kind, where the speed vector ran through the red zone but the tip was in the safe zone. Among the 18 non-conflict scenarios, these six scenarios had the highest false-positive rates: 27% (3 of 11 participants) or 36% (4 of 11 participants). Figure 3 shows the SSD for the three scenarios with a 36% false-positive rate (top row) and three scenarios that yielded a false positive rate of 0% (bottom row). Figure 3 suggests that the high false-positive rates can be explained because participants misunderstood the SSD: The tip is in the safe zone, and hence the aircraft are not in conflict.

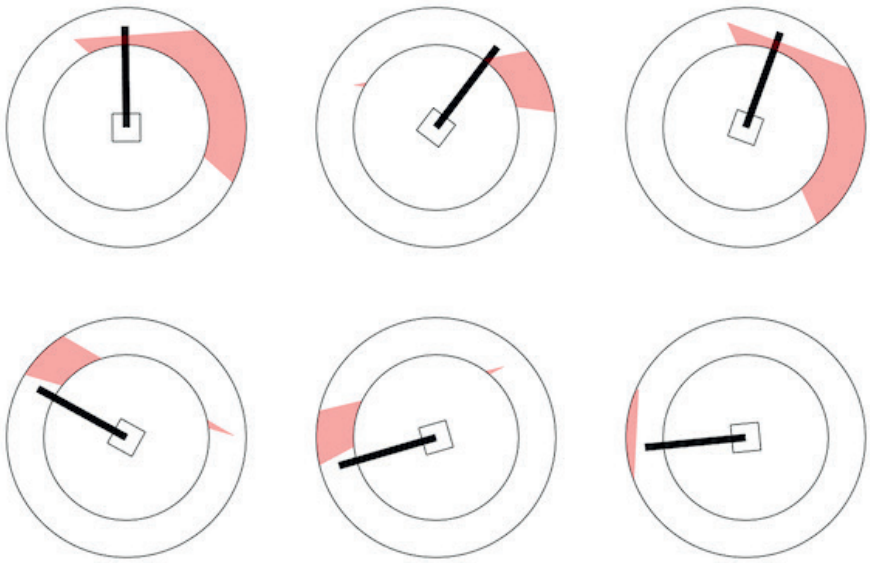


Figure 3. Six selected SSDs in non-conflict scenarios. Top row: SSDs that yielded a high false-positive rate (36%). Bottom row: SSDs that yielded a low false-positive rate (0%). The high false-positive rates may be caused by the fact that the speed vector runs through the red zone.

The mean saccade amplitude was not significantly different between the SSD group and the No-SSD group (Table 2). The mean saccade amplitude was strongly dependent on how far the two aircraft were spaced apart (r = 0.97 for no-SSD participants, r = 0.93 for SSD participants, n = 44 scenarios, see Figure 4). Thus, the saccade amplitude was scenario-specific and not much influenced by the presence of the SSD.

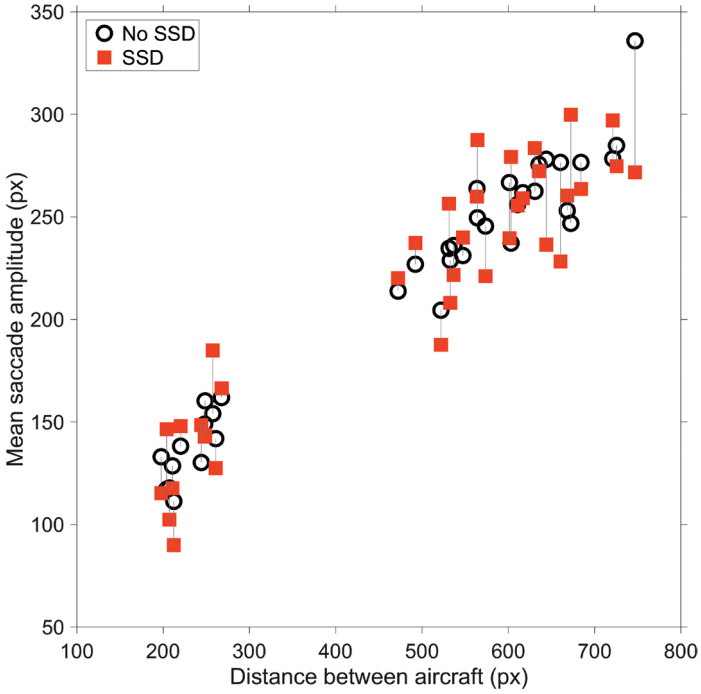


Figure 4. Mean saccade amplitude versus distance between aircraft for the 36 regular scenarios. A vertical line is used to connect the same scenarios.

The participants from the SSD group devoted about twice as much attentional time to Aircraft 1 (which contained the SSD) as compared to participants from the No-SSD group (see Table 2). The long viewing durations of the SSD group at Aircraft 1 came at the expense of attention to other areas of interest, in particular Aircraft 2 and the lines between the Aircraft and the conflict point (Table 2). These findings are illustrated in Figure 5 for one of the scenarios.

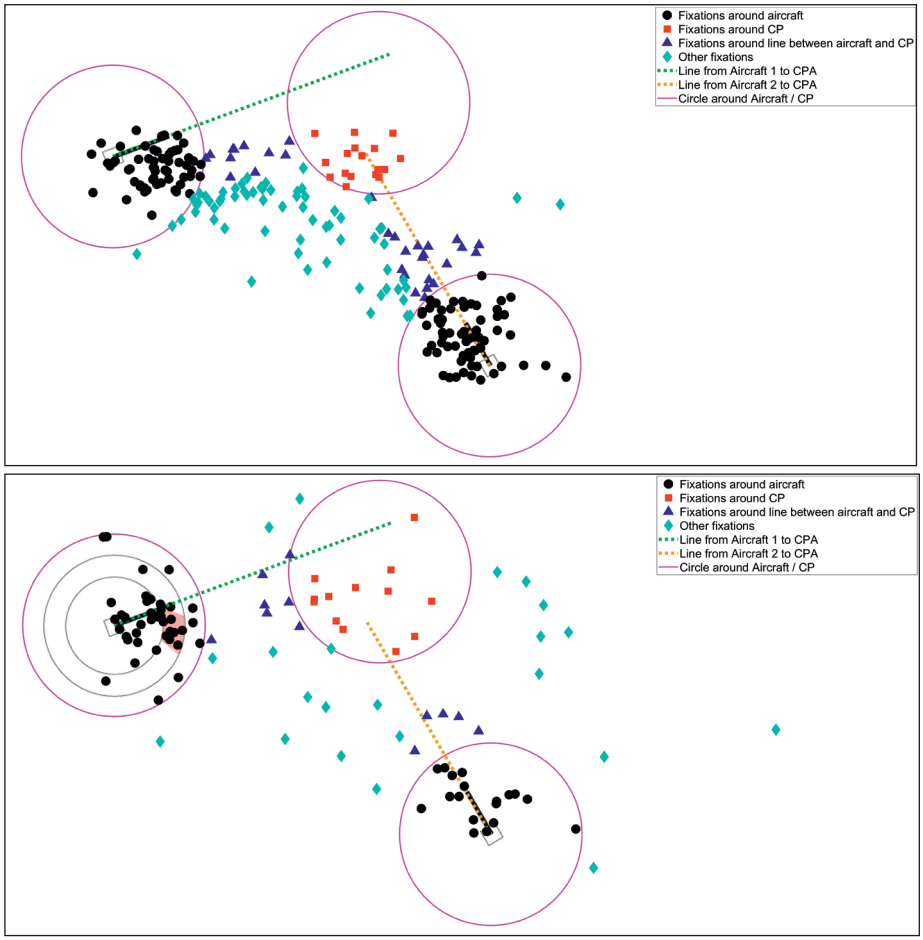


Figure 5. Fixation locations for a non-conflict scenario (Scenario #10, see also Figure 1). The top figure shows 237 fixations of 13 participants in the No-SSD group, and the bottom figure shows 108 fixations of 11 participants in the SSD group. Note that the mean fixation duration of participants in the No-SSD group was shorter (466 ms) as compared to participants in the SSD group (1073 ms). CP = Conflict point, CPA = Closest point of approach.

As a final analysis, we examined the percentage of participants who looked at Aircraft 1 as a function of time during the trial. The results of this analysis, as shown in Figure 6, indicate that Aircraft 1 attracted attention at the start of the trial (i.e., between 0.5 and 1.5 seconds). Furthermore, no clear learning effects can be distinguished from Scenarios 1–18 to Scenarios 23–40.

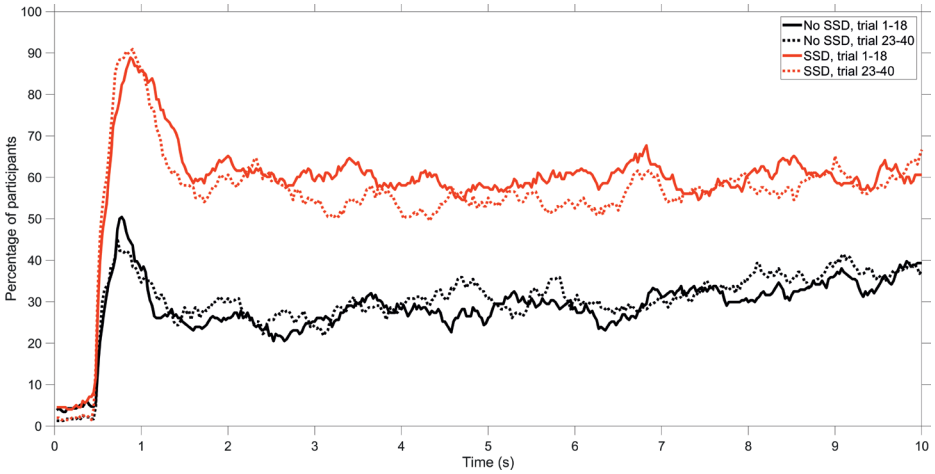


Figure 6. Percentage of participants who looked at the area of interest (AOI) of Aircraft 1. A distinction is made between the no SSD group and the SSD group and between Scenarios 1–18 combined and Scenarios 23–40 combined.

DISCUSSION

This study compared self-reported workload, conflict-detection performance, and distribution of visual attention between novice participants who were supported by a visual aid (the Solution Space Diagram, or SSD) and participants who had to do the task unaided. The results showed that the SSD reduced workload to a substantial extent, from 4.56 to 1.53 on a scale from 0 to 10. Furthermore, with the SSD, participants detected conflicts more accurately and quickly as compared to without the SSD. However, conflict detection with the SSD was imperfect, with a miss rate of 6.6%. There are various possible reasons for this imperfect performance. In particular, participants had only 10 seconds to respond. Secondly, it is possible that some participants did not trust the SSD and therefore rejected its indicated correct solution. Disuse is a well-documented phenomenon in the human-automation literature (e.g., Parasuraman & Riley, 1997; Reagan, Cicchino, & Montalbano, 2019).

The false-positive rates showed no statistically significant differences between the SSD and No-SSD groups. This lack of a significant effect could be due to demand characteristics, where some participants may form a conjecture about the goal of the experiment and adjust their response strategy accordingly. In other words, related to the above explanation about disuse, some participants may have ignored the SSD because they expected that conflicts could still be possible despite the fact the SSD signaled that no conflict was present and was perfectly reliable. Additionally, there are clear indications that some participants misunderstood the SSD. More specifically, some participants did not understand that only the position of the tip of the speed vector is relevant for determining the presence of a conflict. Summarizing, the SSD was shown to

improve conflict-detection performance. However, its effects were not compelling with 6.6% misses and 14.6% false alarms, even though the answer to the conflict-detection task could be readily seen.

We used eye-tracking to measure which elements of the visual scene the participants took into consideration. Results showed that participants from the SSD group allocated more attention to Aircraft 1 (containing the SSD overlay) than participants from the No-SSD group. The attention allocated to the SSD can be interpreted as an epiphenomenon of good task performance or as the cause of good task performance, but also points to dangers in the use of augmented feedback. As augmented feedback comes at the expense of judging the relative positions of relevant aircraft and extrapolating the eye movements towards the conflict point, collisions may go undetected in (the unlikely) case that the SSD would display incorrect information.

The high amount of attention allocated to the SSD could be because participants needed time to extract information from the SSD; fixation duration is an often-used measure of the difficulty of extracting information (Fitts, Jones, & Milton, 1950; Underwood et al., 2011). It could also be that the SSD, because of its salient red color, attracted attention in the absence of other compelling cues in the environment. Besides its appearance, participants themselves may expect the SSD overlay to mean something significant, thereby attracting attention. These notions are consistent with the SEEV model of visual sampling (Wickens & McCarley, 2019), stating that expectancy and visually salient features in the environment are attractors of visual attention.

Limitations

A limitation of our study is that participants were engineering students, not air traffic controllers. However, this limitation may not have severe consequences because the conflict detection task was abstract. The 'aircraft' flew in a two-dimensional plane, and the stimuli did not feature ATC-specific features such as flight labels. Accordingly, our study measured general perceptual skills, and one should not immediately generalize the findings to ATC applications. Second, the task featured static images, as opposed to dynamic videos or interactive simulations. The use of static images may be realistic for conflict detection tasks, as regular radar displays should not be expected to have a high update rate. Third, our study was concerned with conflict detection only. The SSD also facilitates opportunities for conflict resolution, something that was not studied herein. However, we argue that, based on Parasuraman et al.'s (2000) stages of information processing, conflict detection necessarily precedes conflict resolution; it is not possible to resolve a conflict if that conflict is not detected first. Fourth, although the SSD consists of nothing more than two circles, a red polygon, and a vector, it was still misunderstood by a number of participants. Future research could use even simpler displays, such as a salient warning signal or a text message as used in traffic collision avoidance systems

(e.g., “traffic, traffic”). It can be expected that simpler displays reduce the visual load but are also more prone to guidance effects. Winstein, Pohl, and Lewthwaite (1994) hypothesized that “feedback that is relatively more guiding would be expected to have greater detrimental effects on motor learning.” (p. 317).

Recommendations and Implications

The question may arise as to whether augmented displays like the SSD represent what they intend to represent. Borst, Visser, Van Paassen, and Mulder (2019) stated that the SSD “portrays velocity obstacles (or, conflict zones) in speed and heading within the maneuvering envelope of the aircraft under control” (p. 624). An important question is whether people indeed see ‘velocity obstacles’ and not merely ‘lines and a red shape’ without further understanding of the work domain. Future research could use interviews, self-reports, or think-aloud methods to examine what people are phenomenologically perceiving. Furthermore, the perceptual task that was used in our study may not exploit the SSD to its fullest potential. Future research could apply augmented feedback in complex supervisory tasks, where knowledge development is important.

Our work has several implications for display design. Intuitively, it may be expected that display augmentation, whether it be the SSD or any other type of additional visual information, improves performance (Maddox, 1996). Our study showed that augmented feedback from the SSD did improve performance, with the correct detection rate increasing from 79.1% to 93.4% and the false-positive rate decreasing from 17.5% to 14.6%. These improvements may be regarded as underwhelming because the SSD always showed the correct solution, and 100% accuracy should therefore be possible. Clearly, the SSD is no panacea, and participants require more instructions or training about how to use the SSD; such extended training/instructions may be expected to reduce the participants’ error rates caused by the confusing SSD design and may facilitate proper reliance on the SSD. It was also shown that augmented feedback attracts attention at the expense of other elements in the environment at no cost to performance. Finally, the SSD was misunderstood in some scenarios. This finding may have been preventable by providing participants with more explicit instructions about how to interpret the SSD. At the same time, this finding serves as a caution for HMI designers, as it shows that augmented feedback that is designed to increase task performance can actually reduce task performance. Our observations are in line with Yeh, Merlo, Wickens, and Brandenburg (2003), who concluded that extraneous visual elements hinder target detection.

Our findings demonstrate that augmented feedback that is intended to improve conflict detection performance has side effects in the form of attentional demands and misunderstanding. Accordingly, we recommend that augmented feedback should

be used with appropriate caution. Better options might be to offer a more explicit form of decision support that uses minimal visual clutter or to fully automate the decision-making task if the automation is sufficiently reliable.

Supplementary material

Supplementary data and scripts are accessible at: https://doi.org/10.4121/uuid:f689c7d5-c1f4-44e3-9897-581da590ff90

REFERENCES

Bazilinskyy, P., Kyriakidis, M., Dodou, D., & De Winter, J. (2019). When will most cars be able to drive fully automatically? Projections of 18,970 survey respondents. *Transportation Research Part F, 64*, 184-195.

Beed, P., Hawkins, M., & Roller, C. (1991). Moving learners towards independence: the power of scaffolded instruction. *The Reading Teacher, 44*, 648–655.

Bijsterbosch, V. A., Borst, C., Mulder, M., & Van Paassen, M. M. (2016). Ecological interface design: sensor failure diagnosis in air traffic control. *IFAC-PapersOnLine*, *49*, 307–312.

Borst, C., Visser, R. M., Van Paassen, M. M., & Mulder, M. (2019). Exploring short-term training effects of ecological interfaces: a case study in air traffic control. *IEEE Transactions on Human-Machine Systems, 49*, 623–632.

Ehrmanntraut, R. (2004). The potential of speed control. *23rd IEEE Digital Avionics Systems Conference (DASC)*, Salt Lake City, UT, 3.E.3-1–3.E.3-7.

Eisma, Y. B., Cabrall, C. D. D., & De Winter, J. C. F. (2018). Visual sampling processes revisited: Replicating and extending Senders (1983) using modern eye-tracking equipment. *IEEE Transactions on Human Machine Systems, 48*, 526–540.

Eisma, Y. B., Looijestijn, A. E., & De Winter, J. C. F. (2020). Attention distribution while detecting conflicts between converging objects: An eye-tracking study. *Vision*. 4,34

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191.

Fitts, P. M., Jones, R. E., & Milton, J. L. (1950). Eye movements of aircraft pilots during instrument-landing approaches. *Aeronautical Engineering Review*, *9*, 24–29.

Hilburn, B., Westin, C., & Borst, C. (2014). Will controllers accept a machine that thinks like they think? The role of strategic conformance in decision aiding automation. *Aircraft Control Quarterly, 22*, 115–136.

Kaber, D. B., & Endsley, M. R. (2004). The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. *Theoretical Issues in Ergonomics Science*, *5*, 113–153.

Maddox (1996). Critique of a 'Longitudinal study of the effects of Ecological Interface Design on skill acquisition' by Christoffersen, Hunter, and Vicente. *Human Factors, 38*, 542–545.

Mercado-Velasco, G., Mulder, M., & Van Paassen, M. (2010). Air traffic controller decision-making support using the Solution Space Diagram. *AIAA Guidance, Navigation, and Control Conference*, 43, 227–232.

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced 'complacency'. *The International Journal of Aviation Psychology, 3*, 1–23.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans*, *30*, 286–297.

Parasuraman, R., & Riley, V. (1994). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*, 230-253.

Reagan, I. J., Cicchino, J. B., & Montalbano, C. J. (2019). Exploring relationships between observed activation rates and functional attributes of lane departure prevention. *Traffic Injury Prevention, 20,* 424-430.

Schmidt, R. A., & Wulf, G. (1997). Continuous concurrent feedback degrades skill learning: Implications for training and simulation. *Human Factors*, *39*, 509–525.

Sheridan, T. B. (1995). Human centered automation: oxymoron or common sense? *IEEE International Conference on Systems, Man and Cybernetics. Intelligent Systems for the 21st Century*, 823–828.

Sheridan, T. B. (2002). *Humans and automation: System design and research issues*. Human Factors and Ergonomics Society.

Thomas, L. C., & Wickens, C. D. (2006). Display dimensionality, conflict geometry, and time pressure effects on conflict detection and resolution performance using cockpit displays of traffic information. *The International Journal of Aviation Psychology*, *16*, 321–342.

Underwood, G., Crundall, D., & Chapman, P. (2011). Driving simulator validation with hazard perception. *Transportation Research Part F: Traffic Psychology and Behaviour*, *14*, 435–446.

Van Leeuwen, P., De Groot, S., Happee, R., & De Winter, J. (2011). Effects of concurrent continuous visual feedback on learning the lane keeping task. *Proceedings of the Sixth International Driving Symposium on Human Factors in Driver Assessment, Traning and Vehicle Design.* Retrieved from https://ir.uiowa.edu/cgi/viewcontent.cgi?article=1436&context=drivingassessment

Wickens, C. D., & McCarley, J. S. (2019). *Applied attention theory.* CRC press.

Winstein, C. J., Pohl, P. S., & Lewthwaite, R. (1994). Effects of physical guidance and knowledge of results on motor learning: Support for the guidance hypothesis. *Research Quarterly for Exercise and Sport, 65*, 316–323.

Wulf, G., & Shea, C. H. (2004). Understanding the role of augmented feedback: The good, the bad, and the ugly. In A. M.Williams & N. J. Hodges (Eds.), *Skill acquisition in sport: Research, theory and practice* (pp. 121–144). London: Routledge.

Yeh, M., Merlo, J. L., Wickens, C. D., & Brandenburg, D. L. (2003). Head up versus head down: The costs of imprecision, unreliability, and visual clutter on cue effectiveness for display signaling. *Human Factors, 45*, 390–407.

SUPPLEMENTARY MATERIALS

Analysis of Transfer Trials

We included a number of transfer scenarios without the SSD to examine guidance or training effects. Here, guidance effects would be confirmed if participants who previously used the SSD perform *worse* than participants who have never used the SSD, and training effects would be confirmed if participants who previously used the SSD perform *better* than those who never used the SSD.

Scenarios 19–22 contained no conflict, and Scenarios 41–44 contained a conflict. This design allowed us to inspect whether transfer effects faded out with scenario number, that is, whether the transfer effect was stronger during the first transfer scenario (i.e., Scenario 19 for conflicts, and Scenario 41 for non-conflicts) compared to the fourth transfer scenario (i.e., Scenario 22 for conflicts and Scenario 44 for non-conflicts).

During the transfer scenarios, where the participants of the SSD group had to do the task without SSD (scenarios 19–22, 41–44), self-reported difficulty increased back to levels equivalent to the No-SSD group (Table S1 & Figure 2). There were again no significant differences between the two groups for any of the dependent variables (Table S1).

Table S1. Means (standard deviations in parentheses) of dependent variables for the No-SSD group and the SSD group during the transfer scenarios. Also shown are the results for independent-samples t-tests.

	Transfer scenarios			
	No SSD (*n* = 13)	**SSD (*n* = 11)**	***t***	***p***
Difficulty (0 to 10)	5.38 (1.13)	5.91 (1.59)	-0.94	.356
Correct detection (%)	63.5 (16.5)	68.2 (16.2)	-0.70	.488
Detection RT (ms)	5981 (1646)	6268 (1007)	-0.50	.619
False positive (%)	9.6 (19.2)	20.5 (15.1)	-1.52	.144
Saccade amplitude (px)	181 (34)	168 (22)	1.16	.258
Fixation duration (ms)	546 (59)	574 (100)	-0.85	.407
Fixations Aircraft 1 (% of time)	21.4 (5.6)	19.0 (8.3)	0.87	.395
Fixations Aircraft 2 (% of time)	39.8 (7.2)	33.0 (9.5)	2.00	.058
Fixations CP (% of time)	7.1 (4.5)	10.7 (7.3)	-1.45	.162
Fixations lines (% of time)	14.7 (8.7)	16.9 (6.8)	-0.68	.505

Summarizing, our study found no significant transfer effects (Table 2) nor visible experience effects during the regular scenarios (Figures 2 & 6). We showed that when the SSD was withdrawn, participants dropped back to unaided levels of performance and workload. Accordingly, we did not confirm the guidance hypothesis; participants from

the SSD group did not perform significantly worse than participants from the No-SSD group in transfer (i.e., there was no 'negative transfer'). However, our experiment also did not find any evidence for the suitability of SSD as a training tool, as no positive transfer effect was identified. In other words, in the context of our conflict-detection task, the training value of the SSD is debatable. It must be noted that we did not use the SSD as part of a training program that is designed to maximize transfer-of-learning. Such a training program could consist of a scaffolding approach by gradually decreasing the amount of information shown by the SSD (Beed, Hawkins, & Roller, 1991). Thus, the present results reflect the 'plain' learning value of the SSD, not its potential learning value within a dedicated learning environment.

It should be noted that with our sample size of 24, we had limited statistical power. Using G*Power software (Faul et al., 2007), we computed the required effect size for a two-group research design (n_1 = 13, n_2 = 11), assuming a false positive rate of 5% and a statistical power of 80%. The results of this analysis showed that the required effect size (Cohen's *d*) is 1.20, a very strong effect. Hence, our design was powerful enough for detecting strong differences while the SSD was present (see Table 2), but not powerful enough for detecting any small transfer effects that may exist. The lower power can be illustrated using Table S1, where a substantial difference in false positives is depicted (20.5% with SSD, 9.6% without SSD) while this effect is not significant (p = 0.058). Note that a Wilcoxon test for the false positives also yielded no significant effect between the SSD and no-SSD groups (p = 0.065). An issue here is that there were only four non-conflict transfer scenarios (Scenarios 41–44), and therefore the false-positive rate for a participant could be either 0%, 25%, 50%, 75%, or 100%, resulting in high variance between participants. Statistical power is expected to increase when using more participants and more scenarios per participant. For future research into transfer-of-training, we recommend using larger sample sizes.

Another explanation for the lack of transfer effects may lie in the duration of the experiment. The experiment may be too short for any substantial learning effect to occur, as learning within our experimental context (ATC) is a process that often takes days or even weeks.

Our findings are consistent with Borst et al. (2019), who found no statistically significant differences in conflict detection performance during transfer scenarios between participants who had completed a two-day training program using an SSD and participants who had received only instructions. Similarly, Van Leeuwen, De Groot, Happee, and De Winter (2011) found that, in a driving simulator, continuous visual feedback on the lateral position enhanced lane-keeping performance, yet attracted substantial amounts of visual attention and did not yield significant differences with a control group in a retention trial.

In the present study, we used transfer scenarios without SSD; future research could examine cases in which the SSD provides erroneous information, as previously explored by Bijsterbosch et al. (2016). Also, future research could examine whether the SSD benefits the learning of the essentials of conflict detection or whether it supports overreliance that inhibits performance when the SSD is removed. Instead of measuring only eye movements and conflict detection performance, future research could employ interviews, knowledge tests, of think-aloud methods to shed light on participants' cognitive processes.

CHAPTER 7

External Human–Machine Interfaces: The Effect of Display Location on Crossing Intentions and Eye Movements

Eisma, Y. B., Van Bergen, S., Ter Brake, S. M., Hensen, M. T. T., Tempelaar, W. J., & De Winter, J. C. F. (2020). External human-machine interfaces: The effect of display location on crossing intentions and eye movements. *Information, 11*, 13.

ABSTRACT

In the future, automated cars may feature external human–machine interfaces (eHMIs) to communicate relevant information to other road users. However, it is currently unknown where on the car the eHMI should be placed. In this study, 61 participants each viewed 36 animations of cars with eHMIs on either the roof, windscreen, grill, above the wheels, or a projection on the road. The eHMI showed 'Waiting' combined with a walking symbol 1.2 s before the car started to slow down, or 'Driving' while the car continued driving. Participants had to press and hold the spacebar when they felt it safe to cross. Results showed that, averaged over the period when the car approached and slowed down, the roof, windscreen, and grill eHMIs yielded the best performance (i.e., the highest spacebar press time). The projection and wheels eHMIs scored relatively poorly, yet still better than no eHMI. The wheels eHMI received a relatively high percentage of spacebar presses when the car appeared from a corner, a situation in which the roof, windscreen, and grill eHMIs were out of view. Eye-tracking analyses showed that the projection yielded dispersed eye movements, as participants scanned back and forth between the projection and the car. It is concluded that eHMIs should be presented on multiple sides of the car. A projection on the road is visually effortful for pedestrians, as it causes them to divide their attention between the projection and the car itself.

1. INTRODUCTION

In recent years, a substantial number of studies have emerged on external human–machine interfaces (eHMIs) for automated cars. In automated driving, non-verbal communication between the driver and other road users is often impossible, because the driver is not physically present in the driver seat, or because the driver is engaged in a non-driving task. One reason for employing eHMIs would be to substitute the lack of eye-contact and other types of non-verbal communication. A second reason for using eHMIs is to transmit information about the future state of the automated vehicle to other traffic participants. For example, if the path planning software of the automated driving system knows that the vehicle will slow down for an upcoming intersection, the eHMI could accordingly communicate that the vehicle is about to slow down [1]. Thus, eHMIs could communicate information that is not apparent from implicit ways of communication, for example, from the car's acceleration and deceleration.

So far, a number of different eHMIs have been designed. Bazilinskyy et al. [2] provided an overview of 22 eHMI concepts from industry, whereas Rasouli and Tsotsos [3] and Schieben et al. [4] presented a survey of eHMIs that are studied in academic contexts. The eHMIs proposed so far come in a variety of modalities, for example as text and light strips (e.g., as in [5]), as well as in many colours (green, red, cyan; [6,7]). Research has found that text-based eHMIs are regarded as easily understood without learning [1,8], and that text has disadvantages related to legibility from a distance and cross-national interpretability [2]. A scientific consensus regarding the most efficient modality for eHMIs has not been reached so far.

A lesser studied question is where on the car the eHMI should be positioned to attain maximum compliance and decision-making efficiency. A variety of locations for eHMIs have been proposed, including:

1. The windscreen [9–12]
2. The front/grill of the car [1,12–22]
3. The roof of the car [23–26]
4. Near the wheels [27] (also proposed by Colley et al. [28])
5. A projection on the road [8,9,23,29–33]

The positioning of the eHMI is important because pedestrians (and other road users) visually sample the road environment in an intermittent matter [34]. The presented information may be critical to road safety, and should be understood early in time.

From the existing body of literature, an eHMI on the front (grill) or roof of the car seems to be the most frequently used option. These locations are justifiable because

they may easily allow for mounting a communication device. An eHMI that projects a message on the road or an eHMI that is integrated with the windscreen are challenging to manufacture. However, these types of eHMIs hold promise because they can be made larger than regular screen-based eHMIs, enhancing their visibility from a distance. This notion is supported by a study using self-reports by Ackermann et al. [9]. They showed that participants found eHMIs that projected its messages on the windscreen or the ground were regarded as better recognisable than display-based eHMIs. Ackermann et al. [9] pointed out that the relatively large size of the projections was probably an underlying reason for these effects.

Even though research (e.g., [35]) shows that pedestrians and drivers do not make direct eye contact very often, an eye-tracking study by Dey et al. [36] showed that pedestrians tend to look at the windscreen when an approaching car is close by, "likely to seek the intention or information about the situational awareness of the driver" (p. 375). Accordingly, a windscreen-based eHMI may be an attractive location for presenting a message. In the same way, Bazilinskyy et al. [37] found that pedestrians often look at the wheels of parked cars; this provides motivation for using a wheel-based eHMI.

At present, it is unclear which location of the eHMI results in the best-perceived clarity and behavioural compliance among pedestrians. This lack of knowledge impedes the standardisation of eHMI designs. In the present study, we let participants view animated video clips in which automated vehicles drove with an eHMI at one of the five abovementioned locations. Participants were asked to hold the spacebar when they felt safe to cross. Consequently, we examined which type of eHMI resulted in the highest time-percentage of spacebar pressings while the automated vehicle slowed down for the participant. This is a continuous behavioural measurement method that was introduced by De Clercq et al. [1]. Additionally, we used eye-tracking to infer which type of eHMI yields the most concentrated gaze patterns.

A survey of eHMI concepts proposed by the automotive industry indicated that about 50% of the concepts contained a text message of some kind [2]. Research has also shown that the commanding text 'Walk' can be understood without particular training or prior exposure [1,2]. However, the development of commanding-text eHMIs is technologically challenging, because such design requires that the automated vehicle knows for which road user the command is meant. Another disadvantage of commanding texts concerns liability: if an automated vehicle displays 'Walk', and a pedestrian walks onto the road and collides with a third road user, the manufacturer of the automated vehicle may be at fault.

It has further been shown that a light-based eHMI can be perceived as ambiguous without learning [1,8]. For example, it may be unclear whether a green or red light signal

applies to the pedestrian (egocentric perspective) or the automated vehicle (allocentric perspective; [2]).

Our eHMIs consisted of non-commanding text ('Waiting' or 'Driving') combined with an icon. The text on the eHMI was white to avoid the above-mentioned red/green dilemma. We opted for a relatively salient (i.e, large display/projection) and redundant (i.e., text combined with an icon) eHMI to ensure that participants would have no difficulty understanding what the eHMI message means. We do not aim to suggest that a text-based eHMI would be the optimal solution in real traffic. However, because the present study is concerned with examining the effect of eHMI location, we selected an eHMI design that was shown to be effective in previous research in virtual environments.

2. METHODS

2.1. Participants

The participants were 51 males and 10 females. They were all aged between 19 and 27 years (M = 23.0, SD = 1.8). The participants were all students of BSc and MSc studies at the faculty of Mechanical, Maritime and Materials Engineering at the Delft University of Technology, the Netherlands. About half of the participants were recruited based on opportunity sampling within the faculty building, whereas the other half participated for course credit. All participants provided written, informed consent. The research was approved by the TU Delft Human Research Ethics Committee.

2.2. Apparatus

Eye movements were recorded at 2000 Hz using the Eyelink 1000 Plus eye-tracker v5.15 (SR-Research; Ottawa, Ontario, Canada). Participants were asked to place their head in the head support during the entire experiment. The stimuli were shown on a 24-inch BENQ monitor (Taipei, Taiwan) with a resolution of 1920 × 1080 pixels (531 × 298 mm). The refresh rate of the monitor was set at 60 Hz. The distance between the monitor and the head support was 95 cm. Accordingly, the monitor subtended 31 deg and 18 deg horizontal and vertical viewing angles, respectively. The experimental setup is shown in Figure 1.

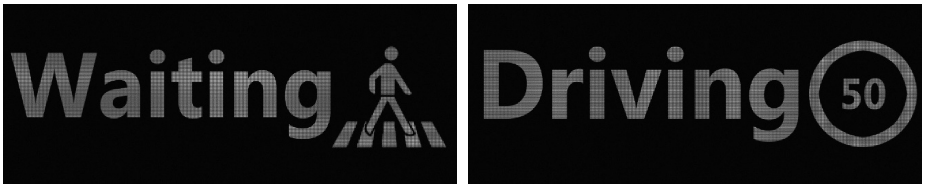
Figure 1. Experimental setup. In the actual experiment, the windows were blinded with aluminium foil.

2.3. Independent Variable

The independent variable was the eHMI type. Six eHMI conditions were used: Roof, Windscreen, Grill, Projection, Wheels, and No eHMI. Figure 2 shows a car that combines all five eHMIs. In the experiment, only one eHMI condition was used at a time. The eHMI could show either 'Waiting' or 'Driving' (Figure 3). The 'Driving' message turned on when the approaching car would not stop for the pedestrian. The 'Waiting' message turned on when the approaching car would stop for the pedestrian.

This study was designed to examine participants' responses when the car was stopping and the eHMI showed 'Waiting'. The responses to the non-stopping vehicles were not analysed herein. The non-stopping vehicles were included to ensure that participants would not start to expect that all cars would stop for them. Note that stopping vehicles had a dominant effect on participants' spacebar-pressing behaviours, whereas no meaningful differences in spacebar-press behaviour between the eHMI conditions occurred for non-stopping vehicles. For example, when the stopping vehicle drove off, it became unsafe to cross, and participants released the spacebar. A non-stopping vehicle that was approaching at that time could not affect spacebar-pressing behaviour because participants already had the spacebar released. We used white text together with a symbol on a black background to achieve the highest possible contrast, because colours (e.g., red and green) already have a meaning, yet this meaning becomes ambiguous when the colour is presented on an approaching vehicle [2].

Figure 2. Car combining all five external human–machine interfaces (eHMIs). In the experiment, the car showed only one eHMI at a time. Here, the car has stopped for the pedestrian. The distance between the centre of the car and the camera (pedestrian) is 7 m longitudinal (i.e., parallel to the direction of the road) and 4.5 m lateral (i.e., perpendicular to the road). The white markings on the road were intended to create a pedestrian crossing on the road, without designated priority to the pedestrian.



(a) **(b)**

Figure 3. (a) Image presented on the eHMI when the approaching car stopped for the pedestrian, (b) Image presented on the eHMI when the approaching car did not stop for the pedestrian.

2.4. Design of the Animated Video Clips

The experiment consisted of 36 non-interactive animated video clips: 6 virtual environments × 6 eHMI conditions. All cars drove at a speed of about 35 km/h unless slowing down for the pedestrian. The videos were 25 s long and played at 60 frames/s. Three environments were used: a straight road, a T-junction and an intersection, with two different preprogrammed traffic behaviours per eHMI. Accordingly, there were six videos per eHMI condition. The lane width was 3.66 m (a standard lane width, e.g., [38]). The camera perspective was from the eyes of a pedestrian waiting to cross the road at a crossing with a traffic island. The field of view of the animation was 80 deg, which ensured that a large part of the environment could be seen (e.g., cars making a right turn, cars driving straight on, and cars making a left turn). In each video, cars were

driving on both lanes. The cars did not contain a driver or passenger. This was done to resemble future driverless vehicles, which may transport goods rather than people.

Within a video, all cars featured the same eHMI type. The eHMI could show one of two messages: If the approaching car passed without slowing down, the eHMI changed from blank to 'Driving' (Figure 3, right). If the approaching car did stop for the participant, the eHMI changed from blank to 'Waiting' (Figure 3, left). The change of state from blank to 'Waiting' occurred when the longitudinal distance between the center of the car and the pedestrian was 23 m. After 1.2 s, when the longitudinal distance had reduced to 11 m, the car started to decelerate to a full stop. The car came to a full stop 2.0 s after the eHMI had switched on, at a longitudinal distance of 7 m between the center of the car and the pedestrian (Figure 2). About 2 s after the car had come to a full stop, the eHMI switched to blank again. About 1.2 s later, the car drove off and passed the participant. These timing and distance parameters yielded a scenario in which cars drove by and stopped in rapid succession. The traffic was not created according to actual traffic data or models of human behaviour.

As stated above, there were six videos per eHMI condition, with each video showing a different traffic environment. The traffic environments were the same for each eHMI, except for a temporal offset (up to 10 s) of the starting moments and corresponding ending moments of the video clips. This offset was included to encourage that participants could not recognise/memorise the behaviour of the cars in the video. In each of the six traffic-environment videos for a particular eHMI condition, one or two of the approaching cars stopped and subsequently drove away. In total, across the six traffic-environment videos per eHMI condition, ten approaching cars stopped for the participant. Details about the video clips and data exclusions are available in the supplementary material (Figures S1–S6).

2.5. Procedure and Task

Participants first read and signed an informed consent form. Next, the eye-tracker was calibrated. Then, participants performed two 10-s training scenarios. These concerned an empty straight road, showing a single car without eHMI; this car approached, stopped and drove off. The participants' task was to press and hold the spacebar whenever they felt it was safe to cross the road. Subsequently, the participants viewed the 36 animated video clips in random order. After each scenario, the participants were asked to rate their perceived clarity with the statement: 'It was clear when I could cross' on a scale from 0 (completely disagree) to 10 (completely agree).

2.6. Dependent Variables

- We calculated the following dependent variables:
- Self-reported clarity on a scale from 0 (completely disagree) to 10 (completely agree).
- Percentage of time that the participant had the spacebar pressed since the moment the eHMI switched to 'Waiting' until 3 s after. A higher percentage indicated a better performance (i.e., indicating when it is safe to cross when it is indeed safe to cross).
- Percentage of time that the participant had the spacebar released since the moment the eHMI switched off before driving away until 3 s after. Again, a higher percentage indicates better performance (i.e., indicating that it is not safe to cross when it is indeed unsafe to cross).
- Gaze spread in pixels. We calculated, for each time sample, the distance between the participant's *x* and *y* gaze coordinates and the mean *x* and *y* gaze coordinates of all participants. The gaze spread is the average distance from the moment the eHMI switched to 'Waiting' until 3 s later.

2.7. Statistical Analyses

The effects of eHMI type on the dependent variables were assessed using a repeated-measures analysis of variance (ANOVA), after averaging the performance scores of the individual vehicle approaches per participant. Significant differences between conditions were assessed with MATLAB's *multcompare* function, using the Tukey–Kramer critical value.

3. RESULTS

3.1. Self-Reported Clarity

Figure 4 shows the results for self-reported clarity per eHMI condition. There was a significant difference between the six eHMI conditions, $F(5,300) = 114.4$, $p < 0.001$, $\eta_p^2 = 0.66$. Pairwise comparisons showed that Roof, Windscreen, and Grill were not significantly different from each other. The mean clarity scores between the other combinations differed significantly.

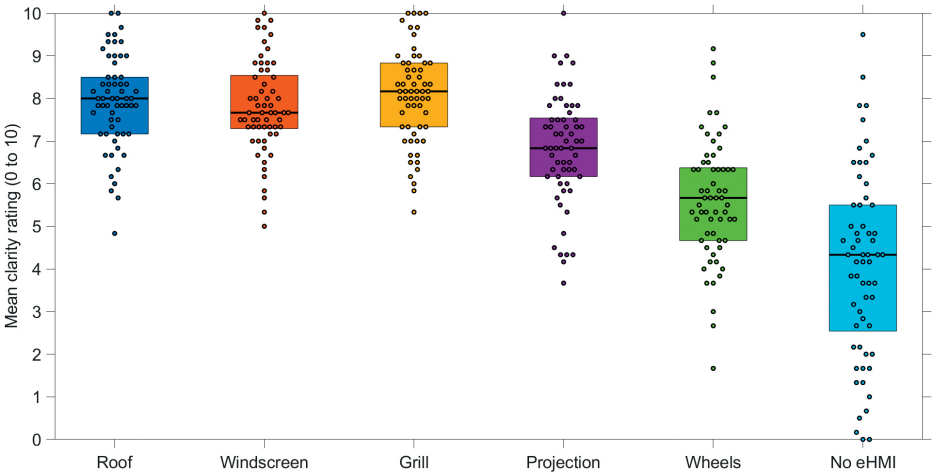


Figure 4. Mean self-reported clarity rating per participant. An average is taken of the scores of six scenarios per participant.

3.2. Performance for Approaching Cars

Figure 5 shows the performance scores, averaged for the nine approaches where the car drove straight on or made a left turn before stopping for the pedestrian. The six eHMI conditions were significantly different from each other, $F(5,300) = 130.1$, $p < 0.001$, $\eta_p^2 = 0.68$. Again, Roof, Windscreen, and Grill were not significantly different from each other, whereas all other combinations differed significantly.

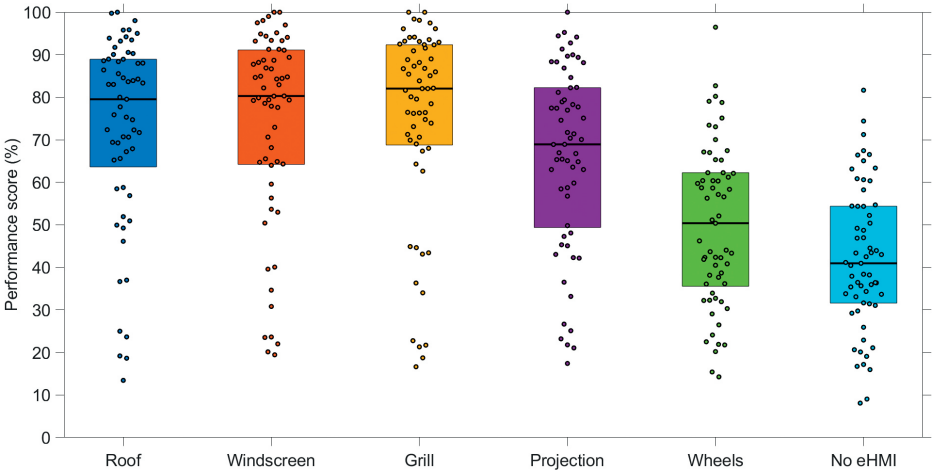


Figure 5. Mean performance score per participant for car approaches. The performance score is defined as the percentage of time that the spacebar was pressed, from the moment the eHMI turned on until 3 s later. The average is taken for the nine approaches where the car drove straight on or made a left turn before stopping for the pedestrian.

Figure 6 illustrates participants' spacebar pressing behaviour as a function of elapsed time since the moment of eHMI onset at t = 0 s. It can be seen that initially (between 0 and 0.5 s), the percentage of participants pressing the spacebar dropped with time, which can be explained by the fact that the approaching car kept getting closer; hence, it became less safe to cross. The Roof, Windscreen, and Grill caused participants to press the spacebar at about 0.5 s since the eHMI turned on. The Projection and especially Wheels triggered a later spacebar-press response, presumably because these eHMIs were poorly visible from a distance; see Figure 7 for an illustration. Figure 6 also shows that for No eHMI, participants only started to press the spacebar once they could detect that the car decelerated (the car decelerated between 1.2 and 2.0 s).

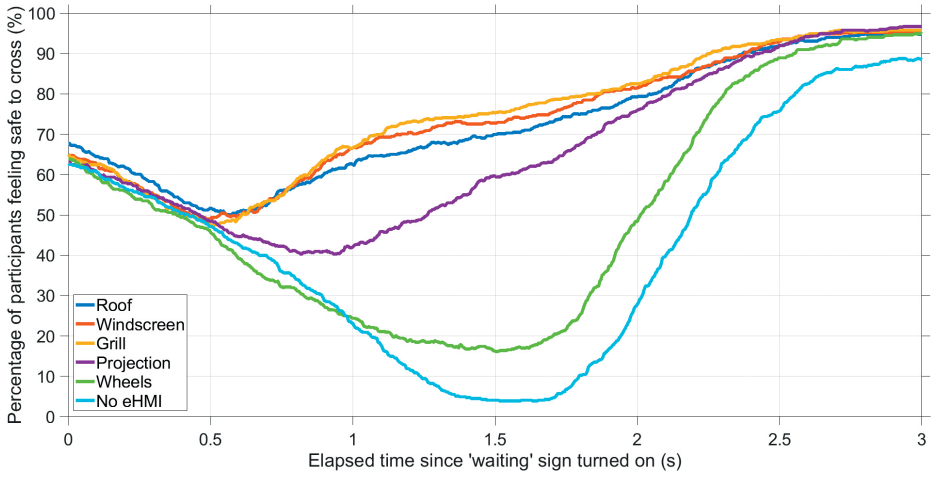


Figure 6. Percentage of participants who pressed the spacebar during car approaches. The average was taken for the nine approaches where the car drove straight on or made a left turn. t = 0 s: the eHMI turns on. t = 2 s: the car has come to a stop.

Figure 7. Screenshot of the animation in a straight approach case with the Projection eHMI. The yellow markers represent the gaze positions of all of the participants. The projection in front of the car is difficult to discern from a distance.

Figure 8 shows the performance score for one selected approach condition: a case where the approaching car made a right turn. Again, the difference in performance scores was significant, $F(5,300) = 10.6$, $p < 0.001$, $\eta_p^2 = 0.15$. All five eHMIs differed significantly from the No eHMI condition, and Wheels differed significantly from Roof and Grill . In other words, in straight and left approach cases, Wheels yielded the lowest performance (Figures 5 and 6), whereas in the right-turn case, Wheels yielded the highest performance (Figure 8).

The high performance for Wheels, and to a lesser extent for Projection, can be explained by the visibility of the sign in the right-turn case (Figure 9). The Roof, Windscreen, and Grill, however, only became visible after the car had made the turn.

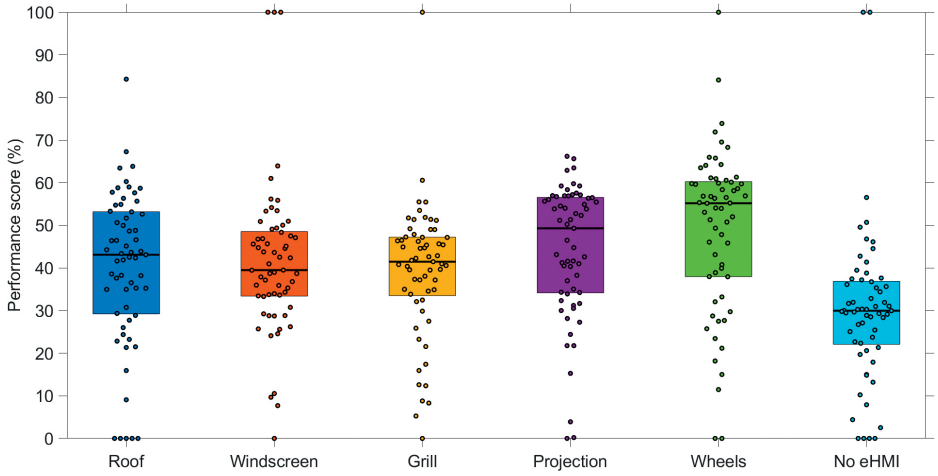


Figure 8. Mean performance score per participant for car approaches where the car made a right turn before stopping for the pedestrian. The performance score is defined as the percentage of time that the spacebar was pressed, from the moment the eHMI turned on until 3 s later.

Figure 9. Screenshot of the animation in the right-turn approach case with the Wheels eHMI. The yellow markers represent the gaze positions of the participants.

The results above showed similar results for self-reported clarity and objective performance. In order to describe the degree of similarity, we averaged the performance scores and clarity scores for all participants per eHMI. The results, shown in Figure 10, reveal a strong association (r = 0.99). In other words, in the aggregate, it appears that clarity and performance are both affected by the same mechanism, which we think is the visibility/readability of the display.

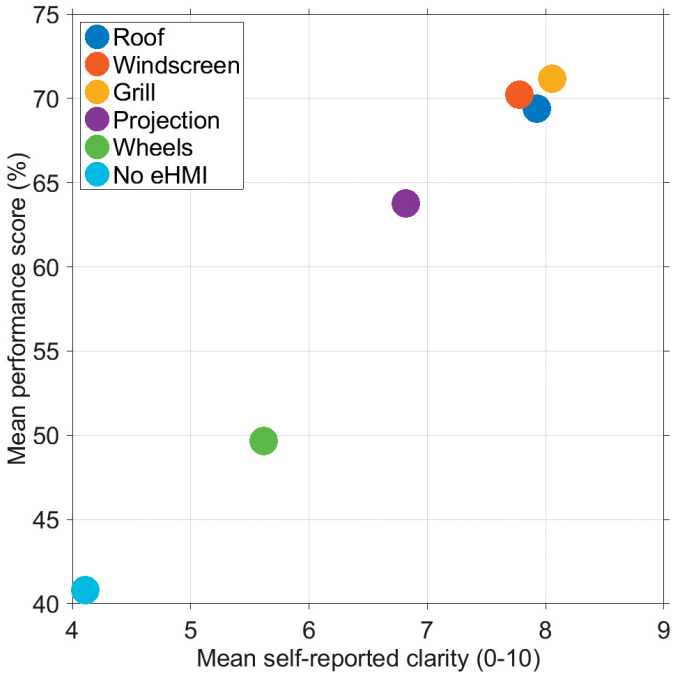


Figure 10. Overall mean self-reported clarity versus overall mean performance score during car approaches. The performance score is defined as the percentage of time that the spacebar was pressed, from the moment the eHMI turns on until 3 s later.

3.3. Eye-Movements for Approaching Cars

A visual inspection of the participants' eye movements indicated that these were often goal-directed, focusing on future interactions. For example, in Figure 11, the majority of participants looked at the approaching car even before the eHMI had turned on; participants did not necessarily look towards the nearest or more salient car. Furthermore, we found that participants' attention distribution was sometimes dispersed (e.g., when multiple cars were visible) and at other times concentrated (e.g., when a relevant car approached the participant, e.g., Figure 9). Herein, we introduce a new measure to describe the degree of gaze dispersion. We defined dispersion as the mean distance from the participants' overall mean gaze coordinate for that particular animated video clip. A dispersion score of, e.g., 200 pixels, means that participants' gaze was, on average, 200 pixels away from the mean fixation gaze position of all participants.

Figure 11. Screenshot of the animation in an intersection scenario. The yellow markers represent the gaze position of the participants.

The results of the gaze dispersion analysis (Figure 12) show that approaching cars attracted attention, as evidenced by low dispersion (< 150 pixels) for the No eHMI condition while the car was approaching (0 to 2 s). The Wheels attracted attention, especially just before coming to a stop (from 1 to 2 s). The Projection, on the other hand, resulted in diversified attention, as illustrated in Figure 13. The Windscreen, on the other hand, yielded in a low gaze dispersion when the car was standing still. The eye-movement dispersion was significantly different between the six eHMI conditions, $F(5,300) = 31.4$, $p < 0.001$, $\eta_p^2 = 0.34$. The Projection yielded a significantly higher dispersion than all five other conditions. The Wheels yielded a significantly lower dispersion than all conditions, except for Windscreen. The Windscreen yielded a significantly lower dispersion than Roof and Projection.

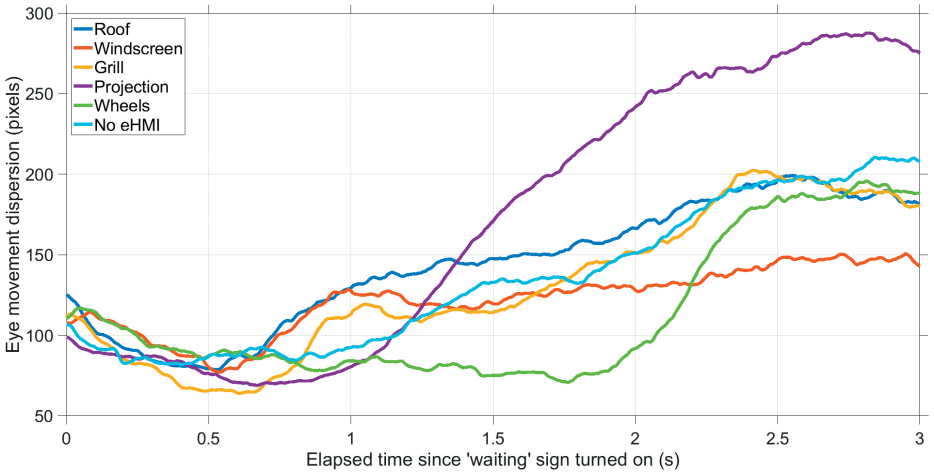


Figure 12. Eye movement dispersion score during car approaches. The average was taken of the nine approaches where the car drove straight on or made a left turn. *t* = 0 s: the eHMI turned on. *t* = 2 s: the car has come to a stop.

Figure 13. Screenshot of the animation in a straight approach scenario with the Projection eHMI. The yellow markers represent the gaze positions of the participants. The Projection results in dispersed eye gaze, with some participants looking at the eHMI on the asphalt and other participants looking at the car.

3.4. Performance for Cars Driving off

So far, we examined only the performance of eHMI for approaching cars. Another relevant aspect of eHMI evaluation is how participants respond after the eHMI switches off before the car drives away. Figure 14 shows that all eHMIs resulted in improved performance compared to No eHMI; that is, participants were more likely to release the spacebar before the car drove off. Initially (at *t* = 0 s), participants using one of the

five eHMIs had the spacebar pressed, because the eHMI displayed 'Waiting' until that point. It took about 0.2 for the first participants to release the spacebar after this eHMI message disappeared. Participants in the No eHMI condition started to release the spacebar only after the car drove off (at 1.4 s), see Figure 14.

An analysis of the performance scores (Figure 15) showed a significant difference between the five eHMI conditions, $F(5,300) = 37.4$, $p < 0.001$, $\eta_p^2 = 0.38$. The No eHMI condition differed significantly from the five other eHMI conditions; there were no significant differences between Roof, Windscreen, Grill, Projection, and Wheels. In other words, participants responded similarly to the eHMI turning off, regardless of the type of eHMI.

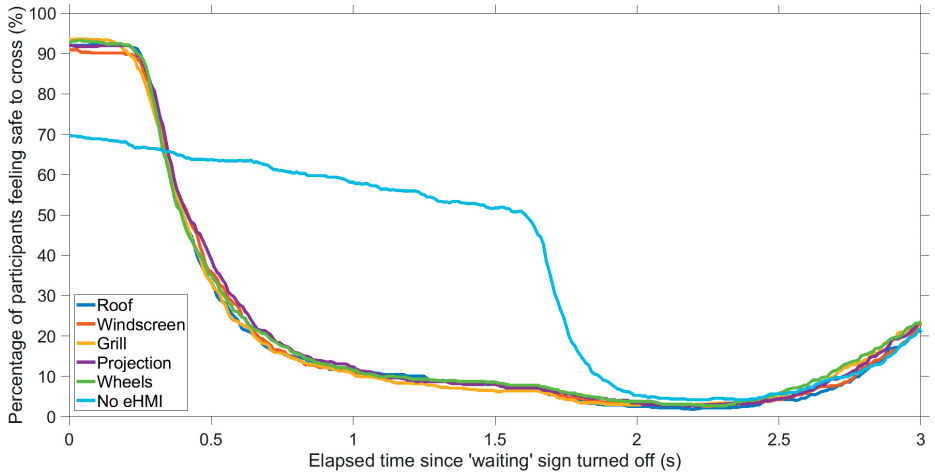


Figure 14. Eye movement dispersion score while the car was driving off. The average is taken of nine times driving off. $t = 0$ s: the eHMI turned off. $t = 1.4$ s: the car started to accelerate.

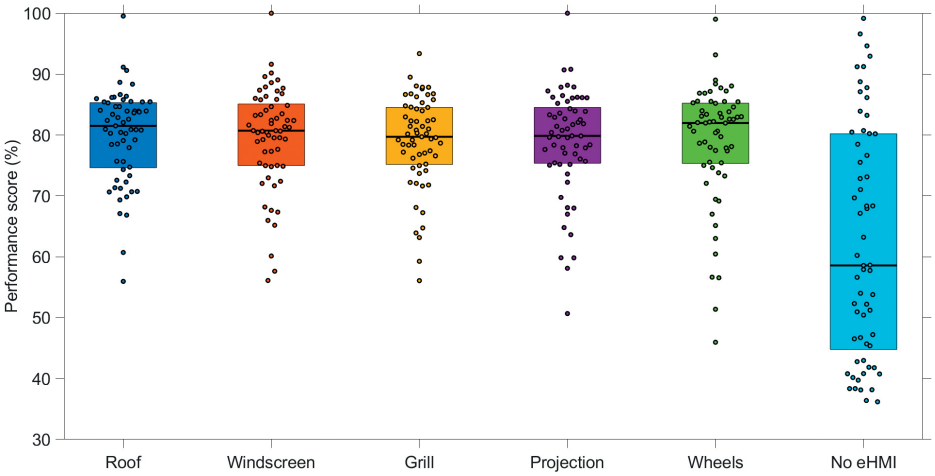


Figure 15. Mean performance score per participant for cases where the car drove off. The performance score is defined as the percentage of time that the spacebar was released, from the moment the eHMI turned off until 3 s later. For each participant, the average is taken of nine times driving off.

4. DISCUSSION

In this study, five eHMI locations, together with a baseline No eHMI condition, were compared in a within-subjects design using a total of 61 participants. The participants viewed animated video clips and were asked to press and hold the spacebar when they thought it was safe to cross, while their eye-movements were recorded using an eye-tracker.

4.1. Performance

The results showed that the Roof, Windscreen, and Grill-based eHMIs yielded the best performance, defined in terms of the pressing time of the spacebar when it was safe to cross. However, this finding did not hold in all scenarios; the eHMI right above the wheel was found to be the best-performing eHMI when the car approached from a corner. In this specific scenario, the eHMIs on the front (Roof, Windscreen, and Grill) were not visible, and therefore failed to communicate their messages to the pedestrian. Together, our findings suggest that eHMIs should be omnidirectional if they are to be applied in traffic scenarios where cars can approach from multiple directions. Vlakveld et al. [26] showed animations of cars with an omnidirectional eHMI on the roof, whereas drive.ai [27] used multiple displays on the car's exterior. Another solution to ensure visibility from all sides is to use a light emitting diode (LED) strip as in Cefkin et al. [39], or LED patterns on the lateral surfaces of the car [40].

The Projection yielded poor spacebar-pressing performance when the car was approaching. This finding can be explained by the poor visibility of the projection at a far distance due to the shallow viewing angle. We do not mean to suggest that our

results generalize to all possible projections. In a virtual reality study, Löcken et al. [31] tested different animations of eHMIs, including a projection which they dubbed F015 (after the name of the concept car presented by Mercedes-Benz USA [33]). Their results showed that the F015 yielded high ratings (5.7 on a scale from 1 to 7) on the User Experience Questionnaire. The concept of Löcken et al. [31] differed from ours, as their projection was highly salient, consisting of a bright green zebra message for the pedestrian. Our findings point to limitations in the use of projections that move with the car, as a projection may not be clear from a distance. We expect that these limitations will be more severe in real traffic. Although technologically feasible (e.g., [41]), it may require powerful lasers to ensure that a projection is visible on the road in daylight. An eHMI on a windscreen may also be technologically challenging to achieve, and may have variable contrast depending on whether or not the eHMI is mounted on a transparent windscreen or whether the windscreen is blinded (in the case of level 5 autonomous vehicles).

For the events where the car was driving away, and the eHMI switched from 'Waiting' to a blank display, all five eHMI locations were found to yield equivalent performance. These findings can be explained because the removal of the message was a salient event, which participants could detect independent of eHMI location or even message content.

Our findings indicate that it is possible to convince users to cross or not to cross before the car slows down or drives away. In other words, all eHMI locations were shown to evoke a more accurate response compared to the No eHMI condition.

4.2. Eye-tracking

The eye-tracking results showed that the Windscreen eHMI yielded a concentrated gaze pattern, which can be explained by the fact that this eHMI is embedded in the centre of the car. This finding is in line with Dey et al. [36], who showed that pedestrians are inclined to look at the windscreen when an oncoming car gets close to the pedestrian. The Wheels eHMI also yielded a concentrated gaze pattern, but only for a brief period of about 1 s before the car came to a full stop. This finding may be explained by the fact that the Wheels eHMI was poorly visible from a distance; when the car came close to the participant, they were inclined to fixate on the eHMI to read its message.

We found that the Projection eHMI yielded a dispersed eye-movement pattern, a finding that can be attributed to the fact that participants looked at the projection and the car itself. These results are consistent with Powelleit et al. [42], who tested a projection in front of the car showing the predicted vehicle trajectory. The results of Powelleit et al. [42] showed that drivers found such a display distracting. Similarly, we see a risk that a projection on the road may result in distraction, where road users may fixate on the

projection on the road at the expense of attention towards the car itself, and therefore may miss relevant implicit cues.

Such results have been found in the use of visual augmented feedback in air traffic control: Eisma et al. [43] found that augmented visual feedback helps to achieve a better task performance, but also has distraction potential.

4.3. Self-Reports

An interesting result was that, in the aggregate, self-reported clarity was strongly associated with objective performance, with a correlation of 0.99. This strong correlation may be due to a single underlying factor, such as the legibility of the display. In other words, the Projection and Wheels eHMIs were hard to read from a distance, as a result of which participants pressed the spacebar late and gave a low clarity rating. The strong correlation between subjective and objective performance is promising for those who examine eHMIs using self-reports (e.g., [8]).

4.4. Limitations and Recommendations

The present study was conducted in rather constrained conditions. We used a computer monitor that offered a physical field of view of 31 deg and a virtual field of view of 80 deg. The 36 videos followed each other in quick succession, and the cars in the videos did not behave according to a realistic traffic flow model. Furthermore, participants were given a straightforward task to press the spacebar when feeling that it was safe to cross.

It would be worthwhile to employ more ecologically-valid methods, such as a virtual reality headset combined with a motion suit [44] or a field test using a Wizard of Oz approach [39]. It remains to be investigated how participants would respond to eHMIs in real traffic, in which situations arise more naturally and in which pedestrians may be in a hurry or lack the concentration to focus on a particular eHMI. We especially recommend testing eHMIs in traffic environments that involve competing visual demands. It is possible that pedestrians in complex traffic rely on peripheral vision without sustained visual attention towards the eHMI [39,45]. Wide fields of view could be achieved using a head-mounted display or surround projections. An advantage of our setup, in which head movement was constrained, is that we were able to measure eye movements with high accuracy.

Our computer monitor had a standard resolution of 1920 × 1080 pixels. The text-based eHMIs may have been hard to read when the virtual car drove at a large distance, especially for participants that suffer from near-sightedness. As discussed above, the Projection eHMI was relatively difficult to perceive just after it has appeared. However, despite the limited display resolution, participants rated the Roof, Windscreen, and Grill eHMIs as clear, with scores of about 8 on a scale from 0 to 10, as shown in Figure

4. Furthermore, our experiment proved to be highly sensitive for detecting differences between eHMIs conditions. To illustrate, 1.5 s after the eHMI turned on, over 70% of the participants pressed the spacebar for the Roof, Windscreen, and Grill eHMIs, compared to only 4% without eHMI. The limitation of display quality also applies to other simulation environments, such as CAVE simulations and head-mounted displays (e.g., [1]). In real traffic, legibility will be affected by other types of visual factors, such as direct sunlight, rain, or smog.

Our simulation did not feature sound. In reality, pedestrians may rely on auditory information to establish the state and relative position of oncoming vehicles. Participants in the simulation were not moving through the virtual environment, and the oncoming car decelerated abruptly while not interacting with the participant. These factors should be improved in future research.

For the present experiment, we selected an eHMI consisting of a non-commanding text message combined with an icon. We do not suggest that this type of eHMI is optimal in real-life applications. Clamann et al. [14] mounted a 32-inch screen on the front of a vehicle, depicting messages that were legible from about 75 m distance. Such large screens, or even multiple screens (see [27]), may not be desirable from an aesthetics and aerodynamics point of view and will require careful system integration. Because display clarity is an essential factor for performance, we recommend that future research examines highly salient eHMI, such as a blinking LED strip.

A final limitation is that the present experiment was conducted using young engineering students, who can be expected to have a relatively high spatial ability [46] and perceptual speed [47]. It remains to be investigated whether older people would be able to intuitively understand eHMIs,such as the ones tested in the present study.

5. CONCLUSIONS

In conclusion, eHMIs on the Grill, Windscreen, and Roof were subjectively regarded as the clearest and evoked the highest rate of compliance for approaching cars. A projection-based eHMI has limitations in the form of poor legibility and participants' visual attention distribution. Based on our results, we recommend that eHMIs should be visible from multiple directions.

REFERENCES

1. De Clercq, G.K.; Dietrich, A.; Núñez Velasco, P.; De Winter, J.C.F.; Happee, R. External human-machine interfaces on automated vehicles: Effects on pedestrian crossing decisions. *Hum. Factors* **2019**. doi:10.1177/0018720819836343.
2. Bazilinskyy, P.; Dodou, D.; De Winter, J.C.F. Survey on eHMI concepts: The effect of text, color, and perspective. *Transp. Res. F Traffic Psychol. Behav.* **2019**, 67, 175–194. doi:10.1016/j.trf.2019.10.013.
3. Rasouli, A.; Tsotsos, J.K. Autonomous vehicles that interact with pedestrians: A survey of theory and practice. *IEEE Trans. Intell. Transp. Syst*. In press. doi:10.1109/TITS.2019.2901817.
4. Schieben, A.; Wilbrink, M.; Kettwich, C.; Madigan, R.; Louw, T.; Merat, N. Designing the interaction of automated vehicles with other traffic participants: Design considerations based on human needs and expectations. *Cognit. Technol. Work* **2019**, *21*, 69–85. doi:10.1007/s10111-018-0521-z.
5. Benderius, O.; Berger, C.; Lundgren, V.M. The best rated human-machine interface design for autonomous vehicles in the 2016 Grand Cooperative Driving Challenge. *IEEE Trans. Intell. Transp. Syst*. **2018**, *19*, 1302–1307. doi:10.1109/TITS.2017.2749970.
6. Zhang, J.; Vinkhuyzen, E.; Cefkin, M. Evaluation of an autonomous vehicle external communication system concept: A survey study. In *Advances in Human Aspects of Transportation. AHFE 2017. Advances in intelligent Systems and Computing*; Stanton, N., Ed.; Springer: Cham, Swizerland, 2017; Volume 597, pp. 650–661. doi:10.1007/978-3-319-60441-1_63.
7. Werner, A. New colours for autonomous driving: An evaluation of chromaticities for the external lighting equipment of autonomous vehicles. *Colour Turn* **2018**, *1*. doi:10.25538/tct.v0i1.692.
8. Fridman, L.; Mehler, B.; Xia, L.; Yang, Y.; Facusse, L.Y.; Reimer, B. To walk or not to walk: Crowdsourced assessment of external vehicle-to-pedestrian displays. *arXiv* **2017**, arXiv:1707.02698.
9. Ackermann, C.; Beggiato, M.; Schubert, S.; Krems, J.F. An experimental study to investigate design and assessment criteria: What is important for communication between pedestrians and automated vehicles? *Appl. Ergon.* **2019**, *75*, 272–282. doi:10.1016/j.apergo.2018.11.002.
10. Nissan. IDS Concept. Available online: https://www.nissan.co.uk/experience-nissan/concept-cars/ids-concept.html (accessed on 2 December 2019).
11. Sweeney, M.; Pilarski, T.; Ross, W.P.; Liu, C. Light Output System for a Self-driving Vehicle. U.S. Patent No. US9902311B2, 2018.
12. Weber, F.; Chadowitz, R.; Schmidt, K.; Messerschmidt, J.; Fuest, T. Crossing the street across the globe: A study on the effects of eHMI on pedestrians in the US, Germany and China. In *HCI in Mobility, Transport, and Automotive Systems*. *HCII 2019. Lecture Notes in Computer Science*; Krömker, H., Ed.; Springer: Cham, Switzerland, 2019; Volume 11596, pp. 515–530. doi:10.1007/978-3-030-22666-4_37.
13. Chang, C.M.; Toda, K.; Igarashi, T.; Miyata, M.; Kobayashi, Y. A video-based study comparing communication modalities between an autonomous car and a pedestrian. In Proceedings of the Adjunct Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Toronto, ON, Canada, 23–25 September 2018; pp. 104–109. doi:10.1145/3239092.3265950.

14. Clamann, M.; Aubert, M.; Cummings, M.L. Evaluation of vehicle-to-pedestrian communication displays for autonomous vehicles. In Proceedings of the Transportation Research Board 96th Annual Meeting, Washington, DC, USA, 8–12 January 2017.

15. Daimler. Autonomous Concept Car Smart Vision EQ Fortwo: Welcome to the Future of Car Sharing. Available online: https://media.daimler.com/marsMediaSite/en/instance/ko.xhtml?oid=29042725 (accessed on 2 December 2019).

16. Joisten, P.; Alexandi, E.; Drews, R.; Klassen, L.; Petersohn, P.; Pick, A.; Abendroth, B. Displaying vehicle driving mode—Effects on pedestrian behavior and perceived safety. In *International Conference on Human Systems Engineering and Design: Future Trends and Applications*; Ahram, T., Karwowski, W., Pickl, S., Taiar, R., Eds.; Springer: Cham, Switzerland, 2019; pp. 250–256. doi:10.1007/978-3-030-27928-8_38.

17. Otherson, I.; Conti-Kufner, A.S.; Dietrich, A.; Maruhn, P.; Bengler, K. Designing for automated vehicle and pedestrian communication: Perspectives on eHMIs from older and younger persons. In *Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2018 Annual Conference*; De Waard, D., Brookhuis, K., Coelho, D., Fairclough, S., Manzey, D., Naumann, A., Onnasch, L., Röttger, S., Toffetti, A., Wiczorek, R., Eds.; 2018; pp. 135–148.

18. Semcon. Who Sees You When the Car Drives Itself? Available online: https://semcon.com/smilingcar (accessed on 2 December 2019).

19. Song, Y.E.; Lehsing, C.; Fuest, T.; Bengler, K. External HMIs and their effect on the interaction between pedestrians and automated vehicles. In *International Conference on Intelligent Human Systems Integration*; Karwowski, W., Ahram, T., Eds.; Springer: Cham, Switzerland, 2018; pp. 13–18. doi:10.1007/978-3-319-73888-8_3.

20. Nuñez Velasco, J.P.; Farah, H.; Van Arem, B.; Hagenzieker, M.P. Studying pedestrians' crossing behavior when interacting with automated vehicles using virtual reality. *Transp. Res. F Traffic Psychol. Behav.* **2019**, *66*, 1–14. doi:10.1016/j.trf.2019.08.015.

21. Stadler, S.; Cornet, H.; Theoto, T.N.; Frenkler, F. A tool, not a toy: Using virtual reality to evaluate the communication between autonomous vehicles and pedestrians. In *Augmented Reality and Virtual Reality*; Tom Dieck, M.C., Jung, T., Eds.; Springer: Cham, Switzerland, 2019; pp. 203–216. doi:10.1007/978-3-030-06246-0_15.

22. Toyota. Concept-i. Available online: https://newsroom.toyota.eu/2018-toyota-concept-i (accessed on 2 December 2019).

23. Deb, S.; Strawderman, L.J.; Carruth, D.W. Should I cross? Evaluating interface options for autonomous vehicle and pedestrian interaction. In Proceedings of the Road, Safety, and Simulation Conference, Iowa City, IA, USA, 14–17 October *2019*.

24. Hensch, A.C.; Neumann, I.; Beggiato, M.; Halama, J.; Krems, J.F. How should automated vehicles communicate?–Effects of a light-based communication approach in a Wizard-of-Oz study. In *International Conference on Applied Human Factors and Ergonomics*; Stanton, N., Ed.; Springer: Cham, Switzerland, 2019; pp. 79–91. doi:10.1007/978-3-030-20503-4_8.

25. Mahadevan, K.; Somanath, S.; Sharlin, E. Communicating awareness and intent in autonomous vehicle-pedestrian interaction. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21–26 April 2018; doi:10.1145/3173574.3174003.

26. Vlakveld, W.; Van der Kint, S.; Hagezieker, M.P. Cyclists' intentions to yield for automated cars at intersections when they have right of way: Results of an experiment using high-quality video animations. Submitted.

27. Drive.ai. The Self-driving Car Is Here. Available online: https://web.archive.org/web/20181025194248/https://www.drive.ai/# (accessed on 2 December 2019).

28. Colley, A.; Häkkilä, J.; Pfleging, B.; Alt, F. A design space for external displays on cars. In Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications Adjunct, Oldenburg, Germany, 24–27 September 2017; pp. 146–151. doi:10.1145/3131726.3131760.

29. Colley, A.; Häkkilä, J.; Forsman, M.T.; Pfleging, B.; Alt, F. Car exterior surface displays: Exploration in a real-world context. In Proceedings of the 7th ACM International Symposium on Pervasive Displays, Munich, Germany, 6–8 June 2018; doi:10.1145/3205873.3205880.

30. Dietrich, A.; Willrodt, J.-H.; Wagner, K.; Bengler, K. Projection-based external human-machine interfaces–Enabling interaction between automated vehicles and pedestrians. In Proceedings of the Driving Simulation Conference Europe, Antibes, France, 5–7 September 2018; pp. 43–50.

31. Löcken, A.; Wintersberger, P.; Frison, A.K.; Riener, A. Investigating user requirements for communication between automated vehicles and vulnerable road users. In Proceedings of the 2019 IEEE Intelligent Vehicles Symposium (IV'19), Paris, France, 9–12 June 2019; pp. 879–884. doi:10.1109/IVS.2019.8814027.

32. Mitsubishi Electric. Mitsubishi Electric Introduces Road-illuminating Directional Indicators. Available online: http://www.mitsubishielectric.com/news/2015/1023.html (accessed on 2 December 2019).

33. Mercedes-Benz USA. Mercedes-Benz F 015 Luxury in Motion. Available online: https://www.youtube.com/watch?v=MaGb3570K1U (accessed on 2 December 2019).

34. Senders, J.W.; Kristofferson, A.B.; Levison, W.H.; Dietrich, C.W.; Ward, J.L. The attentional demand of automobile driving. *Highw. Res. Rec.* **1967**, *195*, 15–33.

35. AlAdawy, D.; Glazer, M.; Terwilliger, J.; Schmidt, H.; Domeyer, J.; Mehler, B.; Fridman, L. Eye contact between pedestrians and drivers. In Proceedings of the Tenth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design, Santa Fe, NM, USA, 24–27 June 2019; pp. 301–307.

36. Dey, D.; Walker, F.; Martens, M.; Terken, J. Gaze patterns in pedestrian interaction with vehicles: Towards effective design of external human-machine interfaces for automated vehicles. In Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Utrecht, The Netherland, 22–25 September 2019; pp. 369–378. doi:10.1145/3342197.3344523.

37. Bazilinskyy, P.; Wesdorp, D.; De Vlam, V.; Hopmans, B.; Visscher, J.; Dodou, D.; De Winter, J.C.F. Visual scanning behaviour on a parking lot. In preparation.

38. Liu, C.; Wang, Z. Effect of narrowing traffic lanes on pavement damage. *Int. J. Pavement Eng.* **2003**, *4*, 177–180. doi:10.1080/1029843042000198586.

39. Cefkin, M.; Zhang, J.; Stayton, E.; Vinkhuyzen, E. Multi-methods research to examine external HMI for highly automated vehicles. In *International Conference on Human-Computer Interaction*; Springer: Cham, Switzerland, 2019; pp. 46–64. doi:10.1007/978-3-030-22666-4_4.

40. Troel-Madec, L.; Alaimo, J.; Boissieux, L.; Chatagnon, S.; Borkoswki, S.; Spalanzani, A.; Vaufreydaz, D. eHMI positioning for autonomous vehicle/pedestrians interaction. In Proceedings of the IHM 2019—31e Conférence Francophone sur l'Interaction Homme-Machine, Grenoble, France, 10–13 December 2019; pp. 1–8.

41. Ineos159challenge The Role of the Car. Available online: https://www.ineos159challenge.com/news/the-role-of-the-car/ (accessed on 2 December 2019).

42. Powelleit, M.; Winkler, S.; Vollrath, M. Cooperation through communication–Using headlight technologies to improve traffic climate. In *Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2018 Annual Conference*; De Waard, D., Brookhuis, K., Coelho, D., Fairclough, S., Manzey, D., Naumann, A., Onnasch, L., Röttger, S., Toffetti, A., Wiczorek, R., Eds.; 2018; pp. 149–160.

43. De Winter, J.C.F., Bazilinskyy, P., Wesdorp, D., De Vlam, V., Hopmans, B., Visscher, J., & Dodou, D. (2021). How do pedestrians distribute their visual attention when walking through a parking garage? An eye-tracking study. Ergonomics, 1–13. https://doi.org/10.1080/00140139.2020.1862310

44. Kooijman, L.; Happee, R.; De Winter, J.C.F. How do eHMIs affect pedestrians' crossing behavior? A study using a head-mounted display combined with a motion suit. *Information* **2019**, *10*, 386. doi:10.3390/info10120386.

45. Moore, D.; Currano, R.; Strack, G.E.; Sirkin, D. The case for implicit external human-machine interfaces for autonomous vehicles. In Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Utrecht, The Netherland, 22–25 September 2019; pp. 295–307. doi:10.1145/3342197.3345320.

46. Wai, J.; Lubinski, D.; Benbow, C.P. Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *J. Educ. Psychol.* **2009**, *101*, 817–835. doi:10.1037%2Fa0016127.

47. Salthouse, T.A. Aging and measures of processing speed. *Biol. Psychol.* **2000**, *54*, 35–54. doi:10.1016/S0301-0511(00)00052-1.

SUPPLEMENTARY MATERIALS

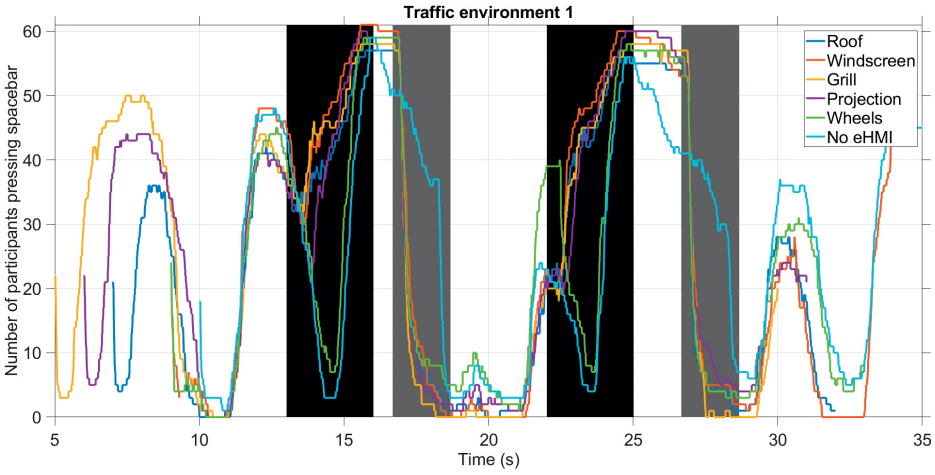


Figure S1. Percentage of participants who pressed the spacebar during the videos of Traffic environment 1. Black background = Car is approaching (3-s period). Gray background = Car is driving off (2-s period).

Comment: The video with the Wheels eHMI accidentally contained a 1-s offset in the behaviour of the car, which can explain the discrepant results in Approach 2. Approach 2 was therefore excluded for the Wheels eHMI analysis.

Traffic environment 1, Approach 1 (straight)

Traffic environment 1, Approach 2 (straight)

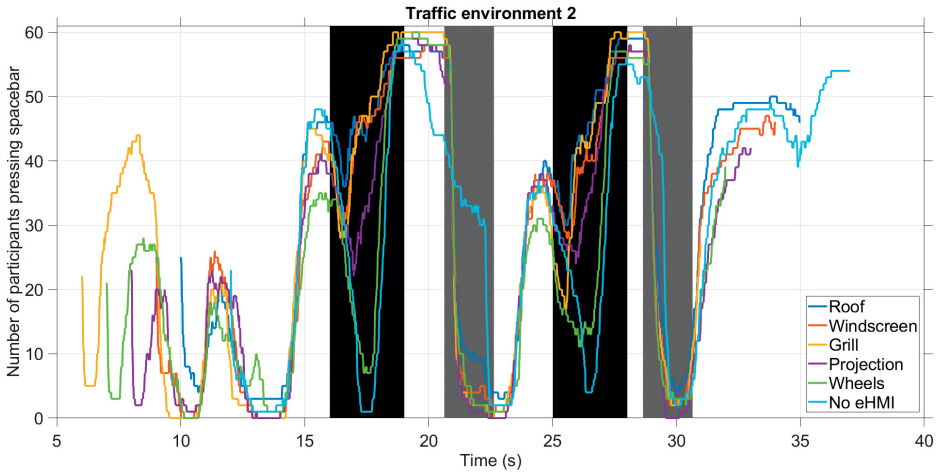


Figure S2. Percentage of participants who pressed the spacebar during the videos of Traffic environment 2. Black background = Car is approaching (3-s period). Gray background = Car is driving off (2-s period).

Traffic environment 2, Approach 1 (straight)

Traffic environment 2, Approach 2 (left turn)

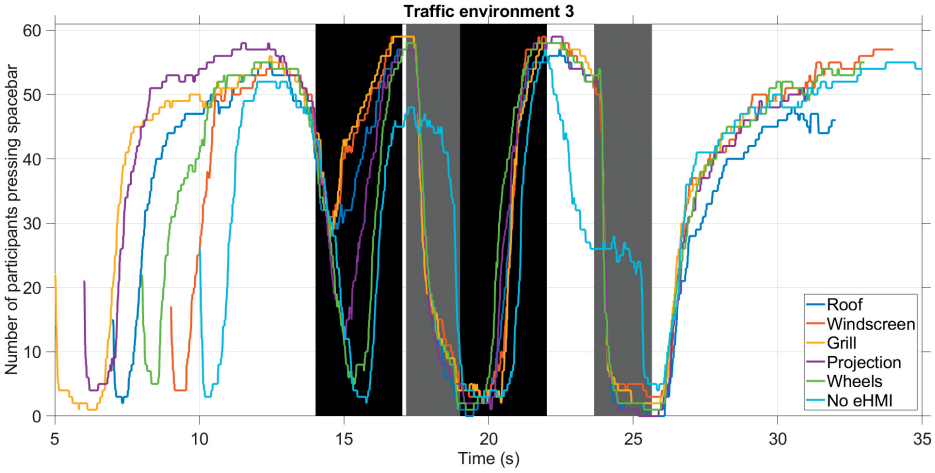


Figure S3. Percentage of participants who pressed the spacebar during the videos of Traffic environment 3. Black background = Car is approaching (3-s period). Gray background = Car is driving off (2-s period).

Traffic environment 3, Approach 1 (straight)

Traffic environment 3, Approach 2 (right turn)

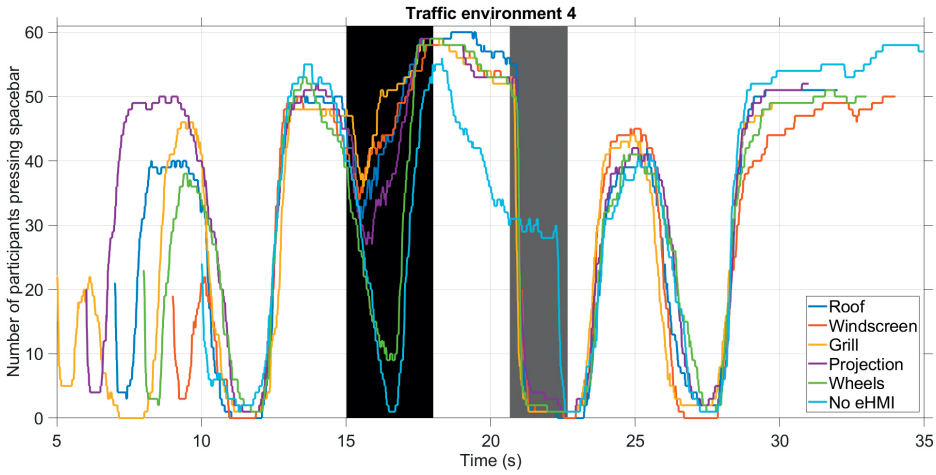


Figure S4. Percentage of participants who pressed the spacebar during the videos of Traffic environment 4. Black background = Car is approaching (3-s period). Gray background = Car is driving off (2-s period).

Traffic environment 4, Approach 1 (straight)

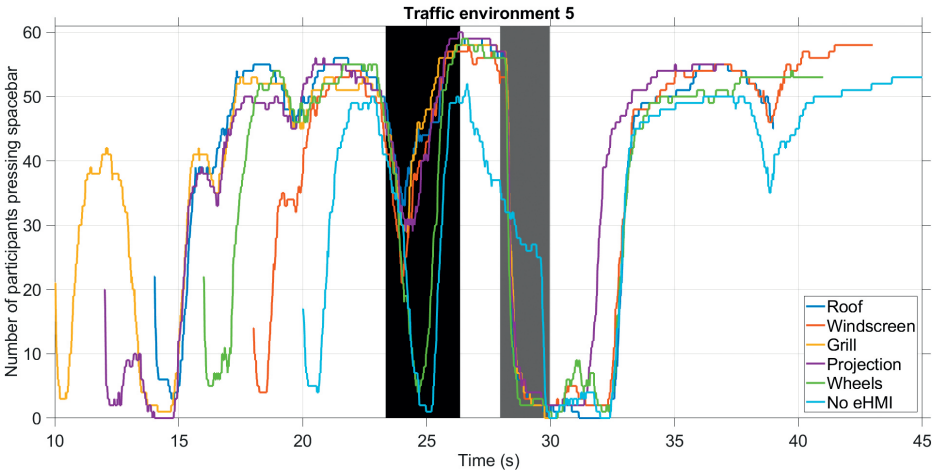


Figure S5. Percentage of participants who pressed the spacebar during the videos of Traffic environment 5. Black background = Car is approaching (3-s period). Gray background = Car is driving off (2-s period).

Comment: The video with the Grill eHMI accidentally lasted 20 s instead of 25 s. The video ends after the period in which the car drove off, so this anomaly does not affect the results. For the Projection eHMI, the behavior of the cars after the car drove off differed from the other videos. This can explain the discrepant results at around 32 s in Figure S5. Again, this anomaly does not affect the results in the paper.

Traffic environment 5, Approach 1 (straight)

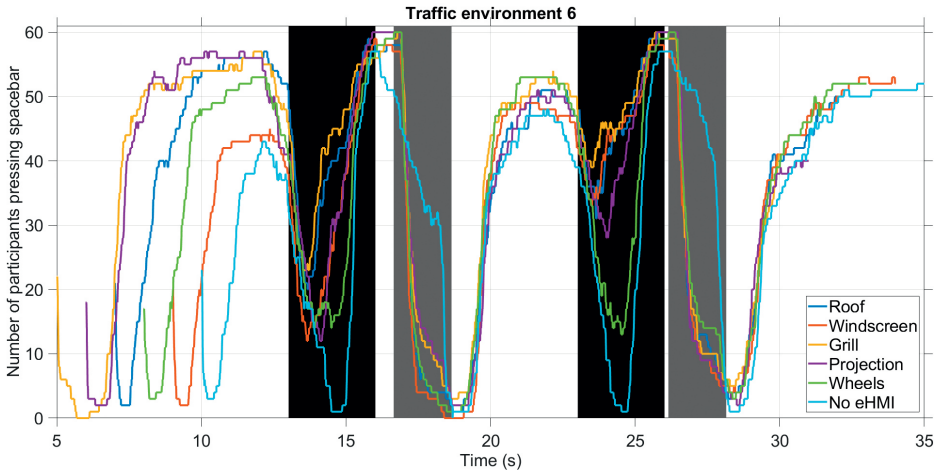


Figure S6. Percentage of participants who pressed the spacebar during the videos of Traffic environment 6. Black background = Car is approaching (3-s period). Gray background = Car is driving off (2-s period).

Comment: Approach 1 for No eHMI and Windscreen were excluded from the analysis because there was not enough time for participants to press the spacebar since the start of the video.

Traffic environment 6, Approach 1 (left turn)

Traffic environment 6, Approach 2 (straight)

CHAPTER 8

External Human-Machine Interfaces: Effects of Message Perspective

Eisma, Y. B., Reiff, A. Kooijman, L., Dodou, D., & De Winter, J. C. F. (in press). External human-machine interfaces: Effects of message perspective. *Transportation Research Part F*, 78, 30–41

ABSTRACT

Future automated vehicles may be equipped with external Human-Machine Interfaces (eHMIs). Currently, little is known about the effect of the perspective of the eHMI message on crossing decisions of pedestrians. We performed an experiment to examine the effects of images depicting eHMI messages of different perspectives (egocentric from the pedestrian's point of view: WALK, DON'T WALK, allocentric: BRAKING, DRIVING, and ambiguous: GO, STOP) on participants' (N = 103) crossing decisions, response times, and eye movements. Considering that crossing the road can be cognitively demanding, we added a memory task in two-thirds of the trials. The results showed that egocentric messages yielded higher subjective clarity ratings than the other messages as well as higher objective clarity scores (i.e., more uniform crossing decisions) and faster response times than the allocentric BRAKING and the ambiguous STOP. When participants were subjected to the memory task, pupil diameter increased, and crossing decisions were reached faster as compared to trials without memory task. Regarding the ambiguous messages, most participants crossed for the GO message and did not cross for the STOP message, which points towards an egocentric perspective taken by the participant. More lengthy text messages (e.g., DON'T WALK) yielded a higher number of saccades but did not cause slower response times. We conclude that pedestrians find egocentric eHMI messages clearer than allocentric ones, and take an egocentric perspective if the message is ambiguous. Our results may have important implications, as the consensus among eHMI researchers appears to be that egocentric text-based eHMIs should not be used in traffic.

INTRODUCTION

With the increasing number of automated vehicles (AVs) on the road, an emerging challenge concerns the interaction between AVs and non-automated road users, such as pedestrians. Traditional ways of communication (e.g., gestures, eye contact) between drivers and pedestrians are likely to disappear, as the AV's 'driver' might be distracted or absent, raising the question as to how communication between AVs and pedestrians should take place (Ackermann, Beggiato, Schubert, & Krems, 2019; Habibovic et al., 2018; Joisten, Freund, & Abendroth, 2020; Stanciu et al., 2018; Sucha, Dostal, & Risser, 2017).

External Human-Machine Interfaces (eHMIs) have been proposed to compensate for the lack of communication between AVs and pedestrians (Lagström & Malmsten Lundgren, 2015). eHMIs appear in several shapes and forms, including text displays, symbolic messages, lights, and projections (Dey et al., 2020a). The design of eHMIs raises concerns regarding the potential ambiguity of the eHMI message. For example, an eHMI in the form of a green braking light on the front of the car could be interpreted in different ways: pedestrians could either think that the meaning of the colour refers to themselves, giving them permission to cross the road (i.e., egocentric perspective) or that the colour refers to the vehicle, indicating its intention to continue driving (i.e., allocentric perspective).

Research in perspective-taking shows that people are inclined to make judgments from their own perspective, whereas adopting another agent's perspective is relatively demanding and error-prone, a phenomenon that has been called egocentric bias or egocentric interference (Ferguson, Apperly, & Cane, 2017; Martin et al., 2019; Surtees & Apperly, 2012). It has been found that the ability to take someone else's perspective decreases with cognitive load (Davis, Conklin, Smith, & Luce, 1996; Lin, Keysar, & Epley, 2010; Roxβnagel, 2000), decreases with time pressure (Epley et al., 2004; Todd, Cameron, & Simpson, 2017), and increases with accuracy incentive and accountability (Epley et al., 2004; Roxβnagel, 2000).

It is presently unknown whether an eHMI should feature an egocentric or allocentric message perspective. We define an egocentric message as a message that the pedestrian can interpret from his/her own perspective. An egocentric message communicates a call to action and addresses the pedestrian. We define an allocentric message as a message which the pedestrian has to interpret from the other agent's (i.e., the AV's) perspective. An allocentric message refers to the action or intention of the AV itself; this means that the pedestrian has to derive the consequences for his/her own actions.

A variety of egocentric and allocentric eHMIs have been proposed in the literature, including the following:

- *Egocentric text-based eHMIs*, such as eHMIs depicting WALK/DON'T WALK (Bazilinskyy, Dodou, & De Winter, 2019; De Clercq, Dietrich, Núñez Velasco, De Winter, & Happee, 2019; Fridman et al., 2019), GO AHEAD (Ackermann et al., 2019; Daimler, 2017), or SAFE TO CROSS (Knight, 2016).
- *Egocentric symbolic eHMIs*, such as eHMIs with a walking pedestrian silhouette (Deb, Strawderman, & Carruth, 2018; Fridman et al., 2019; Hudson, Deb, Carruth, McGinley, & Frey, 2019), a stop sign (Hudson et al., 2019; Urmson, Mahon, Dolgov, & Zhu, 2015), or a raised hand (Fridman et al., 2019; Weber, Chadowitz, Schmidt, Messerschmidt, & Fuest, 2019).
- *Allocentric text-based eHMIs*, including text messages such as AFTER YOU (Nissan, 2015), BRAKING (Deb et al., 2018) and STOPPING (Nissan, 2015).
- *Allocentric symbolic eHMIs*, such as eyes on the car (Chang, Toda, Sakamoto, & Igarashi, 2017), a car with a giving way icon (Weber et al., 2019), or a car depicting it is in automated mode (Joisten et al., 2019).
- *Allocentric light-based eHMIs* depicting the state of the vehicle or of the automated driving system, without specifically addressing pedestrians (Cefkin et al., 2019; Faas, Kao, & Baumann, 2020a; Habibovic et al., 2018; Kaß et al., 2020). It should be noted that some light-based eHMIs use the colour red or green, which the pedestrian should interpret from his/own perspective (e.g., a front brake light in green). In these cases, it can be argued that the light-based eHMIs messages are egocentric rather than allocentric (De Clercq et al., 2019; Petzoldt, Schleinitz, & Banse, 2018; Zhang, Vinkhuyzen, & Cefkin, 2017).

Ackermann et al. (2019) showed that participants prefer instructions/advice about what the pedestrian should do (egocentric messages) over information about the vehicle's status or intention (allocentric messages). Furthermore, Bazilinskyy et al. (2019) found that participants were more inclined to cross in front of an eHMI displaying WALK as compared to an eHMI depicting WILL STOP, again suggesting that egocentric messages are most effective. Of note, egocentric messages are already common in traffic to resolve ambiguities (e.g., a hand gesture to give right of way or traffic signs with a green walking pedestrian). On the other hand, it has been broadly recommended that eHMIs should not offer egocentric messages (i.e., instructions), but should only give information about the state of the vehicle (Faas, Mathis, & Baumann, 2020b; International Organization for Standardization, 2018; Zhang et al., 2017). More specifically, it has been argued that egocentric messages can confuse pedestrians if there are multiple pedestrians in the vicinity of the car (Dietrich, Willrodt, Wagner, & Bengler, 2018) and that egocentric messages may have legal implications in case a pedestrian gets involved in an accident

because he/she complied with the eHMI instructions (Dey et al., 2020a; Tabone et al., 2020).

Study Aim and Approach

This study aimed to investigate the effect of cognitive load and the perspective of the eHMI message on pedestrians' crossing decisions. The experiment was conducted using still images, an approach that resembles previous image-based eHMI experiments (Bazilinskyy et al., 2019; Fridman et al., 2019; Hagenzieker et al., 2020).

We opted for text because, although text requires the participants' visual attention, it appears to be more easily understood than symbolic displays and LED lights (Ackermann et al., 2019; Bazilinskyy et al., 2019; De Clercq et al., 2019). To investigate the effect of message perspective, we selected WALK and DON'T WALK (Bazilinskyy et al., 2019; De Clercq et al., 2019; Fridman et al., 2019; Hudson et al., 2019) for egocentric messages, and BRAKING (Deb et al., 2018) and DRIVING (Eisma et al., 2020) for allocentric messages. In addition, we selected GO (Fridman et al., 2019; Song, Lehsing, Fuest, & Bengler, 2018) and STOP (Fridman et al., 2019; Mercedes-Benz, 2015; Strickland, Yuan, Bai, Weber, & Miucic, 2016; Urmson, Mahon, Dolgov, & Zhu, 2015) as ambiguous messages, defined as messages that can be interpreted either as egocentric or allocentric from a pedestrian's perspective. The ambiguous eHMIs were included to investigate which perspective the participants adopt when the eHMI is not explicit about the perspective to be taken. For example, if most participants indicate not to cross for the message STOP then this would represent an overall egocentric perspective taken by the participants, and if most participants indicate they can cross for the message STOP (as was found by Fridman et al., 2019) this points to an allocentric perspective (i.e., the participants generally assume that the vehicle stops).

Hypotheses

As mentioned above, taking another agent's perspective is cognitively demanding (e.g., Lin et al., 2010). In the context of pedestrian crossing decisions, if the eHMI provides an allocentric message, the pedestrian first needs to interpret what the other agent (i.e., the AV) is going to do, before being able to decide whether he or she can cross the road. In comparison, if the eHMI depicts an egocentric message, the pedestrian could comply with the message directly. Accordingly, we expected that egocentric messages would be regarded as clearer than allocentric messages, where clarity is expressed objectively in terms of the uniformity of crossing decisions (i.e., the extent to which different participants provide the same crossing responses) and subjectively as high clarity ratings. This hypothesis is in line with the works of Ackermann et al. (2019), Bazilinskyy et al. (2019), Clamann, Aubert, and Cummings (2017), De Clercq et al. (2019), and Fridman et al. (2019), who found that the text messages providing advice to the

pedestrian were more preferred, clear, or persuasive than text messages describing the vehicle's state or intent.

For ambiguous message perspectives, the crossing decisions of pedestrians were expected to be less uniform, and response times longer, as compared to the unambiguous (i.e., ego- or allocentric) message perspectives. As people tend to be egocentrically biased, we expected that participants take an egocentric perspective when interpreting ambiguous messages.

In real traffic, pedestrians are likely to integrate information from multiple sources (e.g., vehicles, cyclists, intersection, traffic signs). Furthermore, factors such as visual clutter (Tapiro, Oron-Gilad, & Parmet, 2020), mobile phone use (Bungum, Day, & Henry, 2005; Jiang et al., 2018; Thompson, Rivara, Ayyagari, & Ebel, 2013), or time pressure (Walker, Lanthier, Risko, & Kingstone, 2012) could contribute to additional cognitive load. We added a memory task to the experiment to mimic the cognitive demands that may occur in real traffic. It was expected that with increasing cognitive load, participants would have more difficulty interpreting the meaning of the egocentric and allocentric messages when making a crossing decision, manifested by less uniform decisions and slower response times. Moreover, we expected that for ambiguous messages, participants would make slower and more egocentric decisions (e.g., cross for the message GO, not cross for the message STOP) when cognitive load increases, in line with previous research that suggests that the ability to take someone else's perspective decreases with cognitive load.

During the experiment, we measured eye saccades and pupil diameter using an eye tracker, to make inferences about participants' visual effort and cognitive load, respectively. These measures served as a validation check of the effects of the memory task and allowed us to interpret our findings further.

METHODS

Participants

Hundred and sixty-five MSc students from the Delft University of Technology participated as part of a course. We removed the responses that occurred before the onset of the image (too early) and response times longer than 5000 ms (too late). Participants who responded too early or too late in five or more trials were excluded (N = 62). Accordingly, our final sample consisted of 103 participants (68 males and 35 females), aged between 21 and 29 years (M = 23.3, SD = 2.0). Informed consent was obtained from all participants, and the experiment was approved by the TU Delft Human Research Ethics Committee. All participants were tested individually.

Materials and Equipment

Eye movements were recorded binocularly at a sampling rate of 2000 HZ using an SR-Research EyeLink 1000 Plus eye tracker (see Figure 1). The stimuli were shown on a 24.5-inch BENQ XL2420Z monitor with a resolution of 1920 x 1080 pixels (display area 531 x 298 mm). The distance between the monitor and the table edge was 94 cm. Luminescent lamps on the ceiling lit the room. The participants wore closed-back headphones (Beyerdynamic DT-770 Pro 32 Ohm) to suppress external sounds.

Figure 1. The experimental setup.

Independent Variables

Six eHMIs (Figure 3) were presented. The eHMIs were generated with the online tool LCD Display Screenshot Generator (Avtanski, 2020). We opted for white letters instead of coloured ones to prevent associations with colours that are already used in traffic, such as red and green. Even though the colour cyan is recommended for eHMIs because of its good visibility and for the fact that it is not yet used in traffic (Dey, Habibovic, Pfleging, Martens, & Terken, 2020b; Faas & Baumann, 2019; Werner, 2018), we did not use cyan because it could be misinterpreted as green (Bazilinskyy, Dodou, & De Winter, 2020).

The eHMI concepts were placed on the bumper of a vehicle that contained a driver and a passenger (Figure 2; photo taken from Rodríguez Palmeiro et al., 2018). We included a driver because AVs of SAE levels 1–4 still require human presence (Society of Automotive Engineers, 2019).

Figure 2. One of the six eHMIs used in the experiment. The person in the driver seat provided written consent for the publication of this photograph.

Three independent variables were used. The first independent variable was the perspective of the eHMI message: (1) egocentric (WALK, DON'T WALK), i.e., providing an instruction to the pedestrian, (2) allocentric (DRIVING, BRAKING), i.e., providing information about the state of the vehicle, or (3) ambiguous (STOP, GO), in which case the message perspective could be interpreted either egocentrically or allocentrically. The second independent variable was the yielding intention of the vehicle as conveyed by the eHMI message, that is, whether the vehicle is yielding (WALK, BRAKING) or non-yielding (DON'T WALK, DRIVING). The ambiguous messages STOP and GO were again open to interpretation. The third independent variable was the memory task: we used a forward digit span task, with three levels: 0 digits (baseline), 2 digits (low cognitive load), and 5 digits (high cognitive load). We opted for a maximum of 5 digits based on Miller's law, according to which the number of objects humans can hold in short-term memory is 7 ± 2 (Miller, 1956). After responding to the eHMI stimulus, participants had to type in the digits they remembered. In the baseline condition, participants had to type "0".

Figure 3. The eHMI concepts used in the experiment. Left column: Egocentric messages, Middle column: Allocentric messages, Right column: Ambiguous messages.

Procedure

The experiment consisted of 18 trials (6 eHMIs x 3 memory task levels) that were presented in a random order that was different for each participant. Each of the 18 trials featured a sequence of digits that was the same for all participants. For example, the message BRAKING in the high load condition featured the digits 02809 for all participants. After finishing all trials, participants were asked to rate the clarity of the six eHMI concepts.

In case no memory task was shown, the trial began with a fixation cross at the centre of the screen shown for 5750 ms, followed by a blank screen for 1000 ms. When a memory task was included, the fixation cross was shown for 3750 ms for the low load memory task and 750 ms for the high load memory task, followed by a blank screen for 250 ms and a digit for 750 ms. The blank screen and digit presentation sequence was repeated twice for the low load memory task and five times for the high load memory task. Next, a blank screen appeared for 1000 ms, followed by the statement 'I can cross' shown for 2000 ms, which served to remind the participant about the task. This was followed by a blank screen for 250 ms, a fixation cross for 750 ms, and another blank screen for 250 ms, after which the image with the eHMI was shown until the spacebar is pressed, with a maximum of 5000 ms. Finally, on the last screen of each trial, participants typed in the digits from the memory task. Figure 4 illustrates the presentation sequence of one trial. Note that the time between the onset of the first fixation cross and the onset of the image was identical (10 s) for all trials. A grey background (greyscale level 50% or 127 on a scale from 0 to 255) was used in all cases, except for the eHMI images. The digits and the statement 'I can cross' were presented in a black outline Arial font of 2-pt thickness.

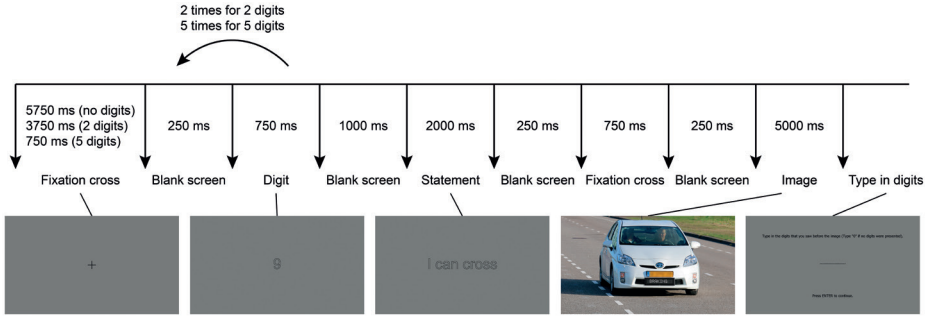


Figure 4. The presentation sequence of one trial.

Participants' task

The participants first read and signed the informed consent form. Participants faced the monitor and adjusted the seat height so that they could comfortably position their head in the head support. They were then presented with an introductory text on the screen that informed them about the contents of the experiment. Note that the experiment described in this paper was the first part of a larger study that included two subsequent unrelated experiments (Eisma & De Winter, 2020). Next, participants completed a standard nine-dot calibration. After the calibration was completed, instructions on the screen informed the participants that they would view images of an AV with textual messages on the bumper and that they had to respond to the statement 'I can cross' by using the L-shift key for 'no' and the R-shift key for 'yes'. These keys were covered with stickers stating 'NO' and 'YES', respectively. Furthermore, they were informed that, for two-thirds of the images, 2 or 5 digits would be shown before the 'I can cross' statement and were explained that they had to remember the digits until after they had responded to the eHMI image. The participants were asked to respond as quickly as possible. One practice trial was performed with a 2-digit memory task and a different eHMI message (WILL STOP) to avoid familiarization.

After the participants had completed all 18 trials, the six images were shown one by one, and the participants rated the clarity of the eHMI message on the vehicle on a scale from 0 (completely disagree) to 10 (completely agree).

Dependent Variables

The following variables were computed:

- *Self-reported clarity.* The participants' response to the statement 'The message on the vehicle is clear' on a scale of 0 (completely disagree) to 10 (completely agree).
- *Objective clarity.* For the messages WALK and BRAKING, the participants were expected to press 'yes' (R-shift), and for DON'T WALK and DRIVING, the participants were expected to press 'no' (L-shift). For the ambiguous eHMI messages, however,

it is undefined whether 'yes' or 'no' constitutes good performance. We have determined a so-called clarity score that allows us to compare the six different conditions in a meaningful way. More specifically, the clarity score was calculated as follows: Objective clarity (%) = 2 x (|percentage of participants pressing 'yes' − 50%|). A score of 100% resembles 'very clear'; that is, participants interpreted the message in a uniform manner. A score of 0% resembles 'very unclear', meaning that 50% of the participants interpreted the message as they could cross the street and 50% as they could not cross the street. Non-responses were excluded from the calculation of objective clarity.

- *Response time.* The response time was measured from the moment when the eHMI image appeared on the screen until the participant pressed the L- or R-shift key.
- *Pupil diameter.* We extracted the participants' pupil diameter from the eye-tracker data. We used pupil diameter as an index of cognitive load, with pupil dilation indicating increased task difficulty (Kahneman & Beatty, 1966). The EyeLink records pupil diameter in arbitrary units. The pupil diameter in millimetres was obtained through a multiplication factor which was based on printed circles of known size.
- *Number of saccades.* Saccades were extracted using a fixation filter previously used by Eisma, Cabrall, and De Winter (2018). The number of saccades during a trial reflects how many eye movements the participants made to reach a decision.

Statistical Analyses

In the literature, there is a growing concern about the use of null hypothesis significance testing, with some voices arguing that *p*-values should be abandoned altogether (Amrhein, Greenland, & McShane, 2019). Consistent with this philosophy, we mostly interpret the data based on point estimates (e.g., means and standard deviations) rather than via *p*-values. However, to assist the reader in identifying which mean values differ significantly from each other, we depict 95% confidence intervals in the figures. For the variables that showed positive correlations between conditions (pupil diameter, response time, self-reported clarity rating, number of saccades), within-subject confidence intervals were computed by first subtracting the participant mean score (for details of this method, see Morey, 2008).

RESULTS

Figure 5 shows the pupil diameter for the three memory task conditions as a function of time. Initially, the pupil diameter declined, which can be explained by the recovery from the previous trial. At t = 10 s, the pupil diameter was higher for the high-load condition (5 digits; M = 4.02 mm, SD = 0.45 mm) as compared to the low-load (2 digits; M = 3.92 mm, SD = 0.45 mm) and baseline conditions (0 digits; M = 3.90 mm, SD = 0.43

mm). The differences between the high-load and baseline condition were statistically significant, as indicated by the nonoverlapping confidence intervals. In other words, the participants experienced cognitive load, as intended. After the presentation of the eHMI at *t* = 10 s, the pupil diameter showed a sharp decline, which can be explained by the pupillary light reflex in response to the increased brightness of the eHMI stimulus as compared to the previous screens.

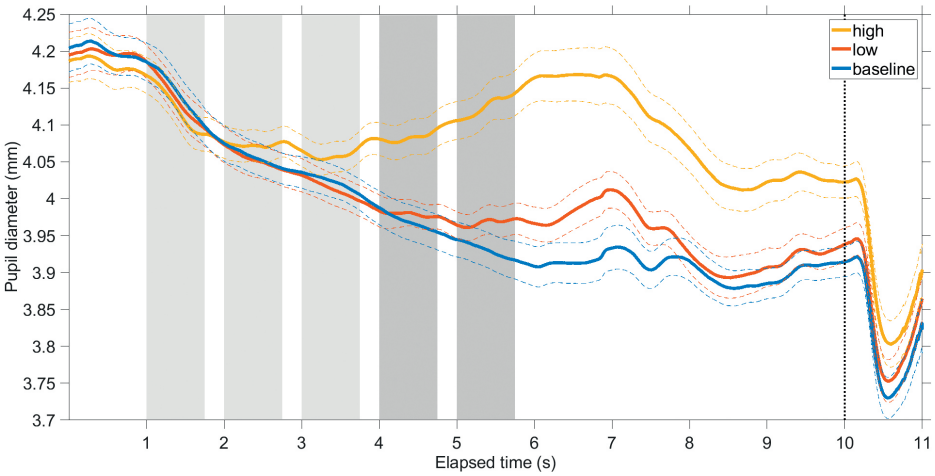


Figure 5. Mean pupil diameter for each memory task condition as a function of time. The grey vertical bands represent the periods when the digits were visible (only the two last darker grey bands for the 2-digit condition and all five grey bands for the 5-digit condition). The black dotted vertical line represents the onset of the stimulus. The dashed lines surrounding the mean pupil diameter are 95% confidence intervals, calculated for each sample point separately.

Figure 6 shows the distribution of responses and the corresponding objective clarity scores per condition. Most participants indicated they could cross for the messages WALK and BRAKING and indicated they could not cross for the messages DON'T WALK and DRIVING, consistent with the intended design. It can also be seen from Figure 6 that the egocentric messages WALK and DON'T WALK yielded the highest objective clarity scores, with the ratio of the number of 'yes' and 'no' responses relative to the total number of responses ('yes' and 'no' combined) being closer to the extremes of the scale (100% and 0%) as compared to the other four conditions. The allocentric BRAKING and the ambiguous STOP yielded the lowest objective clarity scores. For the ambiguous GO, the majority pressed 'yes', and for the ambiguous STOP, the majority pressed 'no', which suggests that the participants took an egocentric perspective.

It was expected that the objective clarity scores would decrease with increasing cognitive load. However, Figure 6 shows that the memory task hardly affected the crossing decisions. For the messages DON'T WALK, clarity scores indeed decreased with

cognitive load, with 99% of participants indicating that they could cross in the baseline condition, compared to 95% for the high-load condition. However, this effect was not large enough to be statistically significant (p = 0.125 according to a two-tailed McNemar test, with n = 97, 4, 0, 1 for no/no, no/yes, yes/no, yes/yes, respectively).

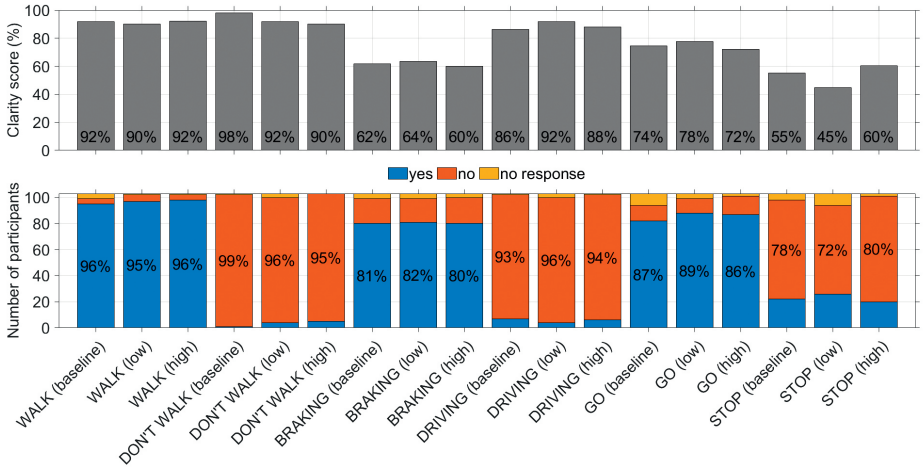


Figure 6. Top: The clarity scores per eHMI condition. Bottom: The corresponding distribution of responses to the statement 'I can cross'. The percentage of 'yes' and 'no' responses is calculated relative to the number of participants who provided a response ('yes' and 'no' combined), excluding non-responses.

Figure 7 shows the corresponding mean response times, with 95% confidence intervals. For the baseline condition (0 digits), the fastest response times were found for egocentric messages, whereas the slowest responses were found for the allocentric BRAKING and the ambiguous STOP. Contrary to our expectations, the memory task reduced the response time compared to the baseline.

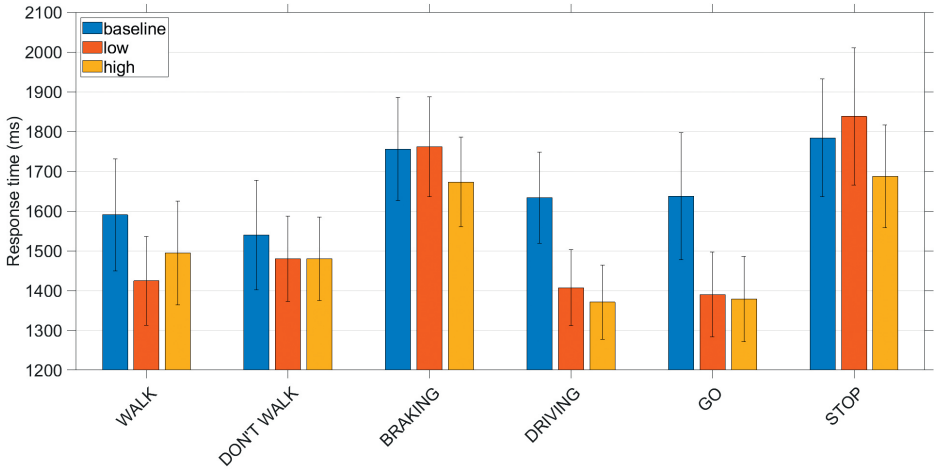


Figure 7. Mean response time. Error bars represent 95% confidence intervals.

Figure 8 shows the cumulative number of saccades, after an offset correction so that the cumulative number of saccades equalled 0 when the stimulus was presented. It can be seen that the message DON'T WALK yielded the largest number of saccades, followed by BRAKING and DRIVING.

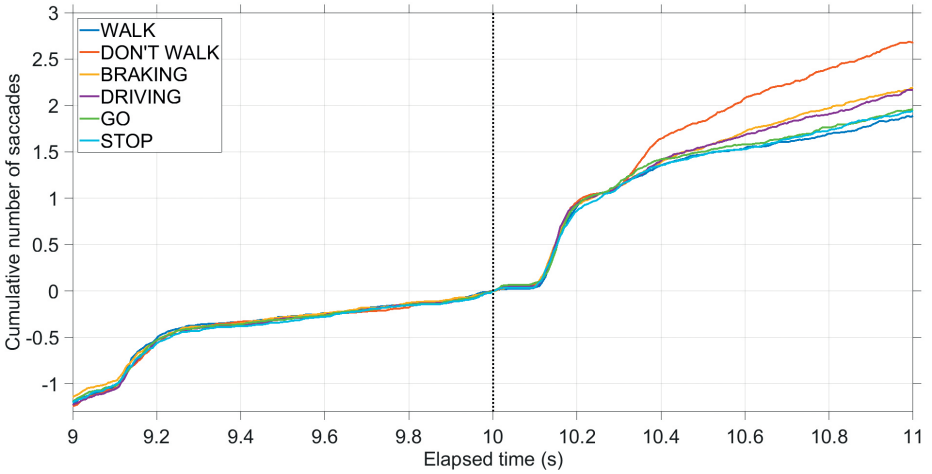


Figure 8. Cumulative number of saccades, after an offset correction so that the value equals zero during the onset of the stimulus. The black dotted vertical line represents the onset of the stimulus.

Figure 9 shows that there was a negative correlation between objective clarity and response time, with the ambiguous STOP and the allocentric BRAKING yielding lower objective clarity and slower responses times than the other conditions (r = -.92, n = 6 conditions). The subjective clarity values were highest for the egocentric messages WALK (M = 8.74, SD = 2.07) and DON'T WALK (M = 9.05, SD = 2.00), followed by the allocentric messages BRAKING (M = 7.19, SD = 2.46) and DRIVING (M = 7.60, SD = 2.43) while the lowest values were obtained for the allocentric GO (M = 6.66, SD = 3.09) and STOP (M = 5.75, SD = 3.52), as illustrated in Figure 9 (middle). Figure 9 (middle) further shows that the mean subjective clarity and objective clarity were strongly positively correlated (r = .87, n = 6 conditions). Figure 9 (right) illustrates that the mean number of saccades at t = 11 s (as shown in Figure 8) correlated positively (r = .92, n = 6 conditions) with the horizontal length of the eHMI message in degrees. Of note, the correlation between the number of saccades and mean response time was near zero (r = -0.13, n = 6 conditions), with the lengthy message DON'T WALK yielding a high number of saccades while being amongst the messages that yielded the fastest response times (Fig. 9, left).

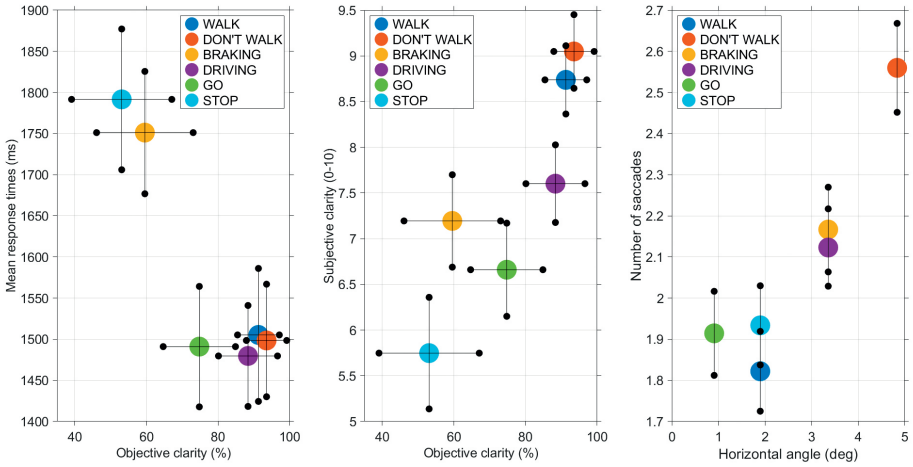


Figure 9. Left: Scatter plot of mean response time and mean objective clarity, averaged across the three memory load conditions. Middle: Scatter plot of mean self-reported clarity and mean objective clarity, averaged across the three memory load conditions. Right: Scatter plot of the mean number of saccades at 11.0 s and the horizontal viewing angle from the leftmost to the rightmost part of the eHMI text message. The error bars represent 95% confidence intervals after averaging the results across the three cognitive load levels per participant.

DISCUSSION

This study aimed to investigate the effect of eHMI message perspective (egocentric, allocentric, ambiguous) as well as cognitive load (baseline, 2 digits, 5 digits) on pedestrians' objective (as computed from their crossing decisions) and subjective (based on their self-reports) clarity levels. The results showed that egocentric messages (WALK and DON'T WALK) yielded higher subjective clarity scores than the allocentric (BRAKING and DRIVING) messages, whereas the egocentric messages yielded higher objective clarity scores than the allocentric BRAKING and the ambiguous STOP. These findings are consistent with our hypotheses, and with previous eHMI studies showing strong pedestrian compliance with egocentric messages (Ackermann et al., 2019; Bazilinskyy et al., 2019; De Clercq et al., 2019; Fridman et al., 2019). A possible explanation for these findings is egocentric bias (e.g., Lin et al., 2010; Todd et al., 2017). That is, participants may naturally be better able to interpret messages that pertain to themselves as compared to messages that pertain to the other (i.e., the AV). Besides egocentric bias, another explanation for the high clarity scores of egocentric messages is that these messages leave little room for misinterpretation (Ackermann et al., 2019; Fridman et al., 2019). In comparison, the allocentric BRAKING might be open to interpretation, as it is unknown whether a braking AV is actually going to a stop for the participant; it can brake for whatever reason. In summary, the effectiveness of egocentric messages may be due to the fact that pedestrians do not have to shift their mental perspective from themselves to the vehicle, but also due to the low ambiguity of these messages.

At the aggregate level, there was a strong positive correlation between objective and subjective clarity (r = .87), a finding which replicates previous research using animated videos that showed a strong correlation between objective performance and subjective clarity for six eHMI conditions (r = .99; Eisma et al., 2020). Of note, the allocentric DRIVING and ambiguous GO resulted in faster responses and higher objective clarity than the allocentric BRAKING and ambiguous STOP. These findings suggest that the time needed to interpret the meaning of the message depends on features that affect perceived clarity, and that, consistent with our argument above, message perspective is not the only factor that affects pedestrian's decision making. We also found that lengthy messages involved a higher number of saccades, presumably reflecting the process of reading the text. However, the negative correlation between the horizontal viewing angle of the text message and the response time suggests that text length was not a contributor to slower responding. In particular, DON'T WALK was the longest text but yielded a fast response.

For the ambiguous messages GO and STOP, most participants were inclined to cross and not cross, respectively. In other words, the majority of the participants interpreted the ambiguous messages from their own point of view. These findings point to an egocentric bias, as in principle, the messages GO and STOP could just as well refer to the AV (i.e., the AV could indicate: 'I stop' or 'I go') or to the pedestrian (i.e., the pedestrian may think: 'I should stop' or 'I can go'). Our findings are consistent with an online study by Vlakveld, Van der Kint, and Hagenzieker (2020), which found that cyclists were more likely to cross when an AV depicted GO as compared to a baseline without eHMI. In an online study by Fridman et al. (2019), on the other hand, most participants interpreted the message STOP allocentrically; that is, participants thought they could cross. The difference with our study was that in Fridman et al., the word STOP was depicted in red, which participants may have associated with a brake light. Because of their inherent ambiguity, we recommend avoiding the words STOP and GO in eHMIs. It is noted that the word STOP is already used in traffic without apparent problems, for example in STOP signs or as a warning not to cross (e.g., a parent may shout STOP to a child when they should not cross the road). The low clarity scores for the message STOP in the present experiment (Figure 9) are likely due to the ambiguity of the perspective to be taken, which arises when this message is attached to an approaching vehicle.

It was expected that the memory task would cause a reduction of objective clarity scores and a slowing of response times. Furthermore, it was expected that for ambiguous messages, cognitive load would contribute to an increase of egocentric crossing decisions (i.e., cross for GO, not cross for STOP). However, contrary to our expectations, crossing decisions were hardly affected by the memory task and participants made *faster* decisions when performing a memory task as compared to not performing the task. An explanation for the faster response times with increasing mental demands

is that participants tried to shed tasks quickly: the faster participants responded, the earlier they could enter their response and the shorter the memory decay (Burke, Allen, & Gonzalez, 2012).

Limitations

This study has a number of limitations. First, we used images to ensure that each participant responded to the same stimulus, without introducing variance in eye movements and decision making caused by vehicle speed and distance. However, the use of images may limit the generalisability of our findings because participants could not make use of vehicle speed to disambiguate the meaning of the eHMI. For example, in reality, a message such as BRAKING may be easier to understand if the vehicle is slowing down at the same time. Previous research suggests that vehicle behaviours are more important than eHMI messages when trying to understand an approaching vehicle's intentions (Clamann et al., 2017; Lee et al., 2020; Li, Dikmen, Hussein, Wang, & Burns, 2018; Moore, Currano, Strack, & Sirkin, 2019). A second limitation is that participants needed to imagine whether they would cross while sitting behind the computer and not having an actual incentive to cross, and not being at risk, which might contribute to response bias. Third, we used a memory task to increase cognitive load, whereas, in real traffic, task load is also determined by visual load and sounds, such as determined by the number of road users. A fourth limitation is that data from about one-third of the participants had to be excluded. Many of these participants misunderstood the task and pressed 'yes' (R-shift) immediately after the statement 'I can cross' was shown.

CONCLUSION AND RECOMMENDATIONS

It is concluded that pedestrians find egocentric messages (WALK, DON'T WALK) clearer than allocentric (BRAKING, DRIVING) and ambiguous (STOP, GO) messages. These findings may be caused by the perspective of the message and associated mental perspective-taking by participants, but may have other causes as well, such as the fact that certain allocentric messages (e.g., BRAKING) are open to multiple interpretations, whereas the messages WALK and DON'T WALK are not. Moreover, it was found that pedestrians take an egocentric perspective if the eHMI message is ambiguous, a finding that provides support for the hypothesis of egocentric bias. Finally, it is concluded that cognitive load in the form of a concurrent memory task reduces response times and that longer messages take longer to read, but do not increase response times. The lengthy message DON'T WALK, for example, yielded a relatively fast response.

Our findings can be placed in the context of recommendations made by experts in eHMI research and design, stating that instructive text-based eHMI messages should not be used in traffic as they could be hard to read from a distance and difficult to understand by people who speak a different language, and might be misleading when

multiple pedestrians have to be addressed at the same time (e.g., Tabone et al., 2020). Our results, however, indicate that egocentric text messages, even two-worded ones, such as DON'T WALK, are responded to quickly and effectively. The present study was conducted in a lab environment with engineering students as participants, and should not be used to make direct inferences about how eHMIs should be deployed in real traffic. In particular, the use of the egocentric message WALK can have negative consequences if a pedestrian decides to cross while this message from the AV was intended for another pedestrian, or when a second vehicle approaches from the opposite direction. Accordingly, the message WALK should perhaps not be used on an eHMI, especially if further research indicates that pedestrians blindly follow up such an instruction even when other indicators indicate that it is not safe to cross. Alternatively, the message SAFE TO CROSS (Bazilinskyy et al., 2019; Knight, 2016) could be used instead of WALK, if the AV (based on its omnidirectional sensor inputs) is confident that it is indeed safe for the pedestrian to cross. In line with the above, we recommend further research into pedestrian compliance and misuse of eHMIs (see also Holländer, Wintersberger, & Butz, 2019; Kaleefathullah et al., in press). Furthermore, it is recommended that future research examines the topic of message perspective for text-based eHMIs versus symbol-based eHMIs, especially in cluttered environments where multiple vehicles are present.

REFERENCES

Ackermann, C., Beggiato, M., Schubert, S., & Krems, J. F. (2019). An experimental study to investigate design and assessment criteria: What is important for communication between pedestrians and automated vehicles? *Applied Ergonomics*, *75*, 272–282. https://doi.org/10.1016/j.apergo.2018.11.002

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. Nature, 567, 305–307. https://doi.org/10.1038/d41586-019-00857-9

Avtanski, A. (2020). LCD display screenshot generator. http://avtanski.net/projects/lcd/

Bazilinskyy, P., Dodou, D., & De Winter, J. (2019). Survey on eHMI concepts: The effect of text, color, and perspective. *Transportation Research Part F: Traffic Psychology and Behaviour*, *67*, 175–194. https://doi.org/10.1016/j.trf.2019.10.013

Bazilinskyy, P., Dodou, D., & De Winter, J. C. F. (2020). External human-machine interfaces: which of 729 colors is best for signaling 'Please (do not) cross'? *IEEE International Conference on Systems, Man and Cybernetics (SMC).*

Bungum, T. J., Day, C., & Henry, L. J. (2005). The association of distraction and caution displayed by pedestrians at a lighted crosswalk. *Journal of Community Health*, *30*, 269–279. https://doi.org/10.1007/s10900-005-3705-4

Burke, M. R., Allen, R. J., & Gonzalez, C. (2012). Eye and hand movements during reconstruction of spatial memory. *Perception*, *41*, 803–818. https://doi.org/10.1068/p7216

Cefkin, M., Zhang, J., Stayton, E., & Vinkhuyzen, E. (2019). Multi-methods research to examine external HMI for highly automated vehicles. *HCI in Mobility, Transport, and Automotive Systems. HCII 2019. Lecture Notes in Computer Science*, *11596*, 46–64. https://doi.org/10.1007/978-3-030-22666-4_4

Chang, C. M., Toda, K., Sakamoto, D., & Igarashi, T. (2017). Eyes on a car: an interface design for communication between an autonomous car and a pedestrian. *Automotive UI '17: Proceedings of the 9th ACM International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 65–73), Oldenburg, Germany. https://doi.org/10.1145/3122986.3122989

Clamann, M., Aubert, M., & Cummings, M. L. (2017). Evaluation of vehicle-to-pedestrian communication displays for autonomous vehicles. *Proceedings of the Transportation Research Board 96th Annual Meeting.* Washington DC.

Daimler. (2017). Autonomous concept car smart vision EQ fortwo: Welcome to the future of car sharing. Retrieved from https://media.daimler.com/marsMediaSite/en/instance/ko/Autonomous-concept-car-smartvision-EQ-fortwo-Welcome-to-the-future-of-car-sharing.xhtml?oid=29042725

Davis, M. H., Conklin, L., Smith, A., & Luce, C. (1996). Effect of perspective taking on the cognitive representation of persons: a merging of self and other. *Journal of Personality and Social Psychology*, *70*, 713–726. https://doi.org/10.1037/0022-3514.70.4.713

Deb, S., Strawderman, L. J., & Carruth, D. W. (2018). Investigating pedestrian suggestions for external features on fully autonomous vehicles: A virtual reality experiment. *Transportation Research Part F: Traffic Psychology and Behaviour*, *59*, 135–149. https://doi.org/10.1016/j.trf.2018.08.016

De Clercq, K., Dietrich, A., Núñez Velasco, J. P., De Winter, J., & Happee, R. (2019). External human-machine interfaces on automated vehicles: Effects on pedestrian crossing decisions. *Human Factors*, *61*, 1353–1370. https://doi.org/10.1177/0018720819836343

Dey, D., Habibovic, A., Löcken, A., Wintersberger, P., Pfleging, B., Riener, A., ... & Terken, J. (2020a). Taming the eHMI jungle: A classification taxonomy to guide, compare, and assess the design principles of automated vehicles' external human-machine interfaces. *Transportation Research Interdisciplinary Perspectives, 7*, 100174. https://doi.org/10.1016/j.trip.2020.100174

Dey, D., Habibovic, A., Pfleging, B., Martens, M., & Terken, J. (2020b). Color and animation preferences for a light band eHMI in interactions between automated vehicles and pedestrians. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3313831.3376325

Dietrich, A., Willrodt, J.-H., Wagner, K., & Bengler, K. (2018). Projection-based external human-machine interfaces – Enabling interaction between automated vehicles and pedestrians. *Proceedings of the Driving Simulation Conference Europe* (pp. 43–50), Antibes, France.

Eisma, Y. B., Cabrall, C. D., & De Winter, J. C. F. (2018). Visual sampling processes revisited: replicating and extending Senders (1983) using modern eye-tracking equipment. *IEEE Transactions on Human-Machine Systems*, *48*, 526–540. https://doi.org/10.1109/THMS.2018.2806200

Eisma, Y. B., & De Winter, J. C. F. (2020). How do people perform an inspection time task? An examination of visual illusions, task experience, and blinking. *Journal of Cognition, 3*, 34. http://doi.org/10.5334/joc.123

Eisma, Y. B., Van Bergen, S., Ter Brake, S. M., Hensen, M. T. T., Tempelaar, W. J., & De Winter, J. C. F. (2020). External human-machine interfaces: The effect of display location on crossing intentions and eye movements. *Information, 11*, 13. https://doi.org/10.3390/info11010013

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology*, *87*, 327–339. https://doi.org/10.1037/0022-3514.87.3.327

Faas, S. M., & Baumann, M. (2019). Light-based external human machine interface: Color evaluation for self-driving vehicle and pedestrian interaction. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *63*, 1232–1236. https://doi.org/10.1177/1071181319631049

Faas, S. M., Kao, A. C., & Baumann, M. (2020a). A longitudinal video study on communicating status and intent for self-driving vehicle–pedestrian interaction. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–14). https://doi.org/10.1145/3313831.3376484

Faas, S. M., Mathis, L. A., & Baumann, M. (2020b). External HMI for self-driving vehicles: which information shall be displayed?. *Transportation Research Part F: Traffic Psychology and Behaviour*, *68*, 171–186. https://doi.org/10.1016/j.trf.2019.12.009

Ferguson, H. J., Apperly, I., & Cane, J. E. (2017). Eye tracking reveals the cost of switching between self and other perspectives in a visual perspective-taking task. *Quarterly Journal of Experimental Psychology, 70*, 1646–1660.

Fridman, L., Mehler, B., Xia, L., Yang, Y., Facusse, L. Y., & Reimer, B. (2019). To walk or not to walk: Crowdsourced assessment of external vehicle-to-pedestrian displays. *Proceedings of Transportation Research Board Annual Meeting*. Washington, DC.

Habibovic, A., Lundgren, V. M., Andersson, J., Klingegård, M., Lagström, T., Sirkka, A., ... & Saluäär, D. (2018). Communicating intent of automated vehicles to pedestrians. *Frontiers in Psychology*, *9*, 1336. https://doi.org/10.3389/fpsyg.2018.01336

Hagenzieker, M. P., Van der Kint, S., Vissers, L., Van Schagen, I. N. G., De Bruin, J., Van Gent, P., & Commandeur, J. J. (2020). Interactions between cyclists and automated vehicles: Results of a photo experiment. *Journal of Transportation Safety & Security, 12*, 94–115. https://doi.org/10.1080/19439962.2019.1591556

Holländer, K., Wintersberger, P., & Butz, A. (2019). Overtrust in external cues of automated vehicles: an experimental investigation. *11th International Conference Automotive User Interfaces*, Utrecht, the Netherlands, 211–222. https://doi.org/10.1145/3342197.3344528

Hudson, C. R., Deb, S., Carruth, D. W., McGinley, J., & Frey, D. (2019). Pedestrian perception of autonomous vehicles with external interacting features. *AHFE 2018. Advances in Intelligent Systems and Computing*, *781*, 33–39. https://doi.org/10.1007/978-3-319-94334-3_5

International Organization for Standardization (2018). ISO/TR 23049: 2018. Road Vehicles - Ergonomic aspects of external visual communication from automated vehicles to other road users. Retrieved from https://www.iso.org/standard/74397.html

Jiang, K., Ling, F., Feng, Z., Ma, C., Kumfer, W., Shao, C., & Wang, K. (2018). Effects of mobile phone distraction on pedestrians' crossing behavior and visual attention allocation at a signalized intersection: An outdoor experimental study. *Accident Analysis & Prevention*, *115*, 170–177. https://doi.org/10.1016/j.aap.2018.03.019

Joisten, P., Alexandi, E., Drews, R., Klassen, L., Petersohn, P., Pick, A., ... & Abendroth, B. (2019). Displaying vehicle driving mode–Effects on pedestrian behavior and perceived safety. *Human Systems Engineering and Design II. IHSED 2019. Advances in Intelligent Systems and Computing*, *1026*, 250–256. https://doi.org/10.1007/978-3-030-27928-8_38

Joisten, P., Freund, A., & Abendroth, B. (2020). Gestaltungsdimensionen der Kommunikation von automatisierten Fahrzeugen und anderen Verkehrsteilnehmenden. *Zeitschrift für Arbeitswissenschaft*, *74*, 132–145. https://doi.org/10.1007/s41449-020-00199-7

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science*, *154*, 1583–1585. https://doi.org/10.1126/science.154.3756.1583

Kaleefathullah, A. A., Merat, N., Lee, Y. M., Eisma, Y. B., Madigan, R., Garcia, J., & De Winter, J. C. F. (in press). External Human-Machine Interfaces can be misleading: An examination of trust development and misuse in a CAVE-based pedestrian simulation environment. *Human Factors*. https://doi.org/10.1177%2F0018720820970751

Kaß, C., Schoch, S., Naujoks, F., Hergeth, S., Keinath, A., Stemmler, T., Keinath, A., & Neukum, A. (2020). Using a bicycle simulator to examine the effects of external HMI on behaviour of vulnerable interaction partners of automated vehicles. *Driving Simulation Conference Europe*, Antibes, France.

Knight, W. (2016). New self-driving car tells pedestrians when it's safe to cross the street. Retrieved from https://www.technologyreview.com/2016/08/30/7287/new-self-driving-car-tells-pedestrians-when-its-safe-to-cross-the-street/

Lagström, T., & Malmsten Lundgren, V. (2015). *AVIP-Autonomous vehicles' interaction with pedestrians – An investigation of pedestrian-driver communication and development of a vehicle external interface* (Master's thesis). Gothenburg, Sweden: Chalmers University of Technology.

Lee, Y. M., Madigan, R., Giles, O., Garach-Morcillo, L., Markkula, G., Fox, C., ... & Dietrich, A. (2020). Road users rarely use explicit communication when interacting in today's traffic: implications for automated vehicles. *Cognition, Technology & Work*. https://doi.org/10.1007/s10111-020-00635-y

Li, Y., Dikmen, M., Hussein, T. G., Wang, Y., & Burns, C. (2018). To cross or not to cross: Urgency-based external warning displays on autonomous vehicles to improve pedestrian crossing safety. *Automotive UI '18: Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 188–197. https://doi.org/10.1145/3239060.3239082

Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, *46*, 551–556. https://doi.org/10.1016/j.jesp.2009.12.019

Martin, A. K., Perceval, G., Davies, I., Su, P., Huang, J., & Meinzer, M. (2019). Visual perspective taking in young and older adults. *Journal of Experimental Psychology: General, 148*, 2006–2026. https://doi.org/10.1037/xge0000584

Mercedes-Benz. (2015). The Mercedes-Benz F 015 Luxury in Motion. Retrieved from https://www.mercedes-benz.com/en/mercedes-benz/innovation/research-vehicle-f-015-luxury-in-motion

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97. https://doi.org/10.1037/h0043158

Moore, D., Currano, R., Strack, G. E., & Sirkin, D. (2019). The case for implicit external human-machine interfaces for autonomous vehicles. *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 295–307). https://doi.org/10.1145/3342197.3345320

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorial in Quantitative Methods for Psychology*, *4*, 61–64. https://doi.org/10.20982/tqmp.04.2.p061

Nissan. (2015). IDS Concept. Retrieved from https://global.nissannews.com/en/releases/release-3fa9beacb4b8c4dcd864768b4800bd67-151028-01-e

Petzoldt, T., Schleinitz, K., & Banse, R. (2018). Potential safety effects of a frontal brake light for motor vehicles. *IET Intelligent Transport Systems, 12*, 449. https://doi.org/10.1049/iet-its.2017.0321

Rodríguez Palmeiro, A., Van der Kint, S., Vissers, L., Farah, H., De Winter, J. C. F., & Hagenzieker, M. (2018). Interaction between pedestrians and automated vehicles: A Wizard of Oz experiment. *Transportation Research Part F: Traffic Psychology and Behaviour*, *58*, 1005–1020. https://doi.org/10.1016/j.trf.2018.07.020

Roxβnagel, C. (2000). Cognitive load and perspective-taking: applying the automatic-controlled distinction to verbal communication. *European Journal of Social Psychology*, *30*, 429–445. https://doi.org/10.1002/(SICI)1099-0992(200005/06)30:3<429::AID-EJSP3>3.0.CO;2-V

Society of Automotive Engineers. (2019). *SAE J3016: automated-driving graphic*. Retrieved from https://www.sae.org/news/2019/01/sae-updates-j3016-automated-driving-graphic

Song, Y. E., Lehsing, C., Fuest, T., & Bengler, K. (2018). External HMIs and their effect on the interaction between pedestrians and automated vehicles. International Conference on Intelligent Human Systems Integration (pp. 13–18). Springer, Cham. https://doi.org/10.1007/978-3-319-73888-8_3

Stanciu, S. C., Eby, D. W., Molnar, L. J., St. Louis, R. M., Zanier, N., & Kostyniuk, L. P. (2018). Pedestrians/bicyclists and autonomous vehicles: how will they communicate? *Transportation Research Record*, *2672*, 58–66. https://doi.org/10.1177/0361198118777091

Strickland, R. D., Yuan, M., Bai, S., Weber, D. W., & Miucic, R. (2016). Vehicle to pedestrian communication system and method (Patent No. 9,421,909). Washington, DC: U.S. Patent and Trademark Office.

Sucha, M., Dostal, D., & Risser, R. (2017). Pedestrian-driver communication and decision strategies at marked crossings. *Accident Analysis & Prevention*, *102*, 41–50. https://doi.org/10.1016/j.aap.2017.02.018

Surtees, A. D., & Apperly, I. A. (2012). Egocentrism and automatic perspective taking in children and adults. *Child Development, 83*, 452-460. https://doi.org/10.1111/j.1467-8624.2011.01730.x

Tabone, W., De Winter, J.C.F., Ackermann, C., Bärgman, J., Baumann, M., Deb, S., Emmenegger, C., Habibovic, A., Hagenzieker, M., Hancock, P. A., Happee, R., Krems, J., Lee, J. D., Martens, M., Merat, N., Norman, D., Sheridan, T. B., & Stanton, N. A. (2021). Vulnerable road users and the coming wave of automated vehicles: Expert perspectives. Transportation Research Interdisciplinary Perspectives, 9, 100293. https://doi.org/10.1016/j.trip.2020.100293

Tapiro, H., Oron-Gilad, T., & Parmet, Y. (2020). Pedestrian distraction: The effects of road environment complexity and age on pedestrian's visual attention and crossing behavior. *Journal of Safety Research*, *72*, 101–109. https://doi.org/10.1016/j.jsr.2019.12.003

Thompson, L. L., Rivara, F. P., Ayyagari, R. C., & Ebel, B. E. (2013). Impact of social and technological distraction on pedestrian crossing behaviour: an observational study. *Injury Prevention*, *19*, 232–237. https://doi.org/10.1136/injuryprev-2012-040601

Todd, A. R., Cameron, C. D., & Simpson, A. J. (2017). Dissociating processes underlying level-1 visual perspective taking in adults. *Cognition*, *159*, 97–101. https://doi.org/10.1016/j.cognition.2016.11.010

Urmson, C. P., Mahon, I. J., Dolgov, D. A., & Zhu, J. (2015). Pedestrian notifications (Patent No. US8954252B1). Washington, DC: U.S. Patent and Trademark Office.

Vlakveld, W., Van der Kint, S., & Hagenzieker, M. P. (2020). Cyclists' intentions to yield for automated cars at intersections when they have right of way: Results of an experiment using high-quality video animations. *Transportation Research Part F: Traffic Psychology and Behaviour, 71*, 288–307. https://doi.org/10.1016/j.trf.2020.04.012

Walker, E. J., Lanthier, S. N., Risko, E. F., & Kingstone, A. (2012). The effects of personal music devices on pedestrian behaviour. *Safety Science*, *50*, 123–128. https://doi.org/10.1016/j.ssci.2011.07.011

Weber, F., Chadowitz, R., Schmidt, K., Messerschmidt, J., & Fuest, T. (2019). Crossing the street across the globe: a study on the effects of eHMI on pedestrians in the US, Germany and China. *HCII 2019. Lecture Notes in Computer Science*, *11596*, 515–530. https://doi.org/10.1007/978-3-030-22666-4_37

Werner, A. (2018). New colours for autonomous driving: An evaluation of chromaticities for the external lighting equipment of autonomous vehicles. *Colour Turn*. https://doi.org/10.25538/TCT.V0I1.692

Zhang, J., Vinkhuyzen, E., & Cefkin, M. (2017). Evaluation of an autonomous vehicle external communication system concept: a survey study. *Advances in Human Factors and Systems Interaction. AHFE 2018. Advances in Intelligent Systems and Computing*, *597*, 650–661. https://doi.org/10.1007/978-3-319-60441-1_63

CHAPTER 9

How Do People Perform an Inspection Time Task? An Examination of Visual Illusions, Task Experience, and Blinking

Eisma, Y. B., & De Winter, J. C. F. (2020). How do people perform an Inspection Time task? An examination of illusions, learning, and blinking. *Journal of Cognition, 3,* 34.

ABSTRACT

In the inspection time (IT) paradigm, participants view two lines of unequal length (called the Pi-figure) for a short exposure time, and then judge which of the two lines was longer. Early research has interpreted IT as a simple index of mental speed, which does not involve motor activity. However, more recent studies have associated IT with higher-level cognitive mechanisms, including focused attention, task experience, and the strategic use of visual illusions. The extent to which these factors affect IT is still a source of debate. We used an eyetracker to capture participants' (N = 147) visual attention while performing IT trials. Results showed that blinking was time-dependent, with participants blinking less when the Pi-figure was visible as compared to before and after. Blinking during the presentation of the Pi-figure correlated negatively with response accuracy. Also, participants who reported seeing a brightness illusion had a higher response accuracy than those who did not. The first experiment was repeated with new participants (N = 159), enhanced task instructions, and the inclusion of practice trials. Results showed substantially improved response accuracy compared to the first experiment, and no significant difference in response accuracy between those who did and did not report illusions. IT response accuracy correlated modestly (r = 0.18) with performance on a short Raven's advanced progressive matrices task. In conclusion, performance at the IT task is affected by task familiarity and involves motor activity in the form of blinking. Visual illusions may be an epiphenomenon of understanding the IT task.

INTRODUCTION

Inspection Time (IT) is defined as "the time required by a subject to make a single observation or inspection of the sensory input on which a discrimination of relative magnitude is based" (Vickers & Smith, 1986), or less formally, "the minimum time required to tell the difference between two perceptually different things" (Irwin, 1984, p. 47). In the standard IT paradigm, the participant views two vertical lines of different lengths, connected by a horizontal line at the top. The participants are exposed to this so-called Pi-figure for a brief time and subsequently have to indicate which of the two lines, the left or the right one, was the longer one. IT is then defined as the exposure time for which participants achieved a threshold accuracy level (e.g., 90% correct). Alternatively, performance at an IT task can be defined as the percentage of trials that were answered correctly (e.g., Ritchie, Bates, Der, Starr, & Deary, 2013).

In a meta-analysis of 92 studies, Grudnik and Kranzler (2001) estimated the mean IT-IQ correlation at –0.30, or –0.51 after correcting for attenuation and range restriction. Early research has theorized that IT scores are an index of mental speed, and therefore a valid indication of psychometric intelligence (Brand, 1981; Brand & Deary, 1982). Jensen (2006) argued that IT is a sensitive index of the "speed of perceptual intake" (p. 84) because participants merely have to determine the difference in a visual stimulus, with no need for providing an immediate motor response as would be the case in, for example, reaction time tasks. Elsewhere, Kranzler and Jensen (1989) mentioned: "IT, the only index of mental speed that does not involve either motor (output) components or executive cognitive processes (metaprocesses), is held to tap individual differences in the 'speed of apprehension,' the quickness of the brain to react to external stimuli prior to any conscious thought." (pp. 329–330). Similarly, Gregory, Nettelbeck, Howard, and Wilson (2008) argued that IT could be used as a biomarker for cognitive decline because an IT task, unlike a reaction time task, is free from psychomotor confounding and does not involve a speed-accuracy trade-off. According to Deary (2000), IT is the simplest possible index that shows a strong correlation ($|r| > 0.3$) with IQ.

Stankov (2004) lamented that "even today some writings on IT, particularly by the ardent supporters of biological interpretations of intelligence, sound like the author(s) believe it is synonymous with intelligence" (p. 351). The current consensus, however, is that to equate a performance measure (IT) with mental speed would be an oversimplification, and that the mechanisms of association between IT and intelligence differences are far from fully understood (Deary, 2001). Structural equation models of Johnson and Deary (2011), for example, suggest that IT may have no unique relationship to general intelligence, and that IT is just one of the elementary cognitive tasks in the broader structure of cognitive ability.

One possible reason for IT not being a pure index of mental speed and intelligence is that IT may be affected by higher-level cognitive mechanisms. According to Deary and Stough (1996), the possibility of IT being a consequence of intelligence differences would represent a validity threat of the IT paradigm: "the inspection-time measure would lose much of its apparent attraction for intelligence researchers, because it would become just another task that clever people perform well" (p. 603).

Several types of cognitive mechanisms for performing IT tasks have been reported in the literature. First, about 50% or more of participants report using cues from visual illusions to perform better at the IT task (e.g., Alexander & Mackenzie, 1992; Chaiken & Young, 1993; Egan & Deary, 1992; Egan, 1994; and see Grudnik & Kranzler, 2001 for a meta-analysis). The two most commonly reported illusions in the IT task are the apparent movement illusion, where people perceive the shorter of the two lines of the Pi-figure to grow as it is overlaid by the mask, and the flash brightness illusion, where people see a bright flash originating from the shorter of the two lines (Alexander & Mackenzie, 1992; Simpson & Deary, 1997). A number of studies have shown that participants who report using illusions perform substantially better at the IT task than nonusers (Egan & Deary, 1992; Egan, 1994; Mackenzie & Bingham, 1985; Mackenzie & Cumming, 1986). Various authors have examined whether different types of masks prevent the perception of illusions and accordingly increase the validity of the IT measurement (e.g., Evans & Nettelbeck, 1993; Stough, Bates, Mangan, & Colrain, 2001), or whether the mask is needed at all (for further discussion, see Egan, 1993).

The second type of cognitive mechanism concerns the effects of experience and practice. It is well established that performance on neuropsychological tests, such as tests of memory and attention, improve with experience (e.g., Seibel, 1963; Sullivan et al., 2017). For IT tasks as well, it has been found that participants perform better if they are re-tested (Anderson, Reid, & Nelson, 2001; Blotenberg & Schmidt-Atzert, 2019; Bors, Stokes, Forrin, & Hodder, 1999; Larson, Saccuzzo, & Brown, 1994; Nettelbeck & Vita, 1992). These findings call into question the notion that IT represents an unmalleable mental quality, and suggest that IT is under the influence of executive functioning or associative mechanisms. For example, participants may come to understand how to perform the task through self-monitoring of past and current performance (Nettelbeck, 2001). The fact that the IT task is susceptible to task familiarity effects has been implicitly acknowledged by the inclusion of familiarization trials (e.g., Bors et al., 1999; Deary et al., 2004; Duan, Dan, & Shi, 2013). However, it is unknown how IT performance improves with practice.

The third type of cognitive mechanism concerns attention (e.g., Bors et al., 1999; Nettelbeck, 2001). It has been found that persons with higher IQ exhibit shorter fixations in visual search tasks than normal-IQ persons, suggesting a link between attention

and IQ (e.g., Sargezeh, Ayatollahi, & Daliri, 2019). Levy (1992) presented the attention hypothesis, which states that IT reflects how well a participant sustains attention to the task. White (1996) pointed out that the micro-deployment of attention is a possible validity threat of the hypothesis that IT is a fundamental task of visual discrimination. For example, IT performance may be better for participants who are visually attentive during the task-critical moments, that is, when the Pi-figure stimulus is visible. The attention hypothesis relates to research which indicates that lapses in attention are related to working memory, executive control, and intelligence (Adam & deBettencourt, 2019; Larson & Alderton, 1990; Oberauer, 2019; Unsworth, Redick, Lakey, & Young, 2010).

So far, attention levels during IT tasks have been measured in indirect ways. Egan and Deary (1992) let participants perform an IT task concurrently with a mental arithmetic task. The participants who reported illusions for the single IT task did not report them in the dual-task condition. Noteworthily, participants who reported illusions in the single-task condition had an IT in the dual-task condition that was shorter than that of participants who did not perceive illusions in the single-task condition, suggesting that

illusions are merely a by-product of good performance. Anderson (1989) let participants perform the IT task in a self- or forced-paced manner, under the assumption that self-pacing reduces distraction. In addition, he applied a fixed versus random period between the end of one IT trial and the beginning of the next and argued that attentional processes would be inhibited if the period were random. Results confirmed expectations that the random period in the forced-paced condition yielded the longest ITs. Hutton, Wilding, and Hudson (1997) let children perform a test battery that measured attentional abilities and subsequently controlled for attention by including IT together with the attention scores in a regression analysis for predicting IQ. Results showed that IT was a statistically significant predictor of IQ even when the attention scores were included in the regression model. The above studies indicate that IT is associated with attention, but do not elucidate the mechanisms of focused attention while performing an IT task.

Several studies have used physiological measures to examine how participants attend to the IT task. Nettelbeck, Robson, Walwyn, Downing, and Jones (1986) presented five experiments in which the eye-movements of low- and normal-IQ participants were measured while performing an IT task. The results showed that the low-IQ participants were prone to distraction before target onset. For example, in one of their experiments, the average number of off-target eye movements was 16.1 out of 240 trials for low-IQ participants, whereas the normal-IQ participants exhibited none. In the same experiment, the number of blinks averaged at 10% and 5% of trials for low- and normal-IQ participants, respectively, an effect that was not statistically

significant. Further research on the role of attention was performed by Deary et al. (2004), who let participants perform an IT task in combination with fMRI. They found elevated activity in select regions of the brain, which they interpreted as effort-related processes and cognitive processes related to attention, working memory, imagery, and vision. Caryl (1994) found significant correlations between IT and ERPs 100 to 200 ms after the stimulus onset and noted that "perhaps ability to focus attention is the fundamental difference between individuals in this task rather than a difference in speed of perceptual intake" (p. 43). More recently, Hill et al. (2011) let a high- and low-IQ group perform an IT task while measuring their ERPs. Based on the larger N1 response for the high-IQ group, they suggested that the link between IT and IQ can be attributed to individual differences in spatial attention. The studies of Deary et al. and Hill et al. indicate that IT is a complex task in which attentional processes play a role. However, so far, it is still unknown *how* people attend to the IT task.

In summary, the validity of IT as an index of 'low-level' mental speed has been questioned from the perspective of three cognitive mechanisms: (1) self-reported visual illusions, (2) experience effects, and (3) attention. The extent to which these factors affect IT is presently a source of debate. This study attempts to extend the findings of previous research by examining how illusions relate to IT performance, how participants improve their IT performance as a function of trial number, and how attention relates to IT performance. Attention in this study was operationalized as 'not blinking', consistent with Johns, Crowley, Chapman, Tucker, and Hocking (2009), who found that reaction times were impaired when blinks occurred during the stimulus onset.

METHODS

Participants

One hundred forty-eight MSc engineering students participated. The data for one participant were not recorded correctly. The remaining 147 participants were 45 females and 102 males with a mean age of 23.33 years (*SD* = 2.13). Twenty-two participants used contact lenses and 13 used glasses. The number of participants who wore glasses during the experiment was smaller than 13, as an undocumented number of participants were encouraged to remove their glasses to enhance the quality of the eye-tracking data.

Apparatus

Movements of the right eye were recorded at 2000 Hz using the SR Research EyeLink 1000 Plus. Participants were asked to keep their head in the head support during the entire experiment.

The visual stimuli were presented using a computer running 'SR Research Experiment Builder' (version 1.10.1386) on a 64-bit Windows 7 Professional operating system. The computer contained an Intel Core i7-4790K Processor (@ 4.00 GHz) and NVIDIA GeForce

GTX 970 graphics card. The stimuli were shown on a 24.5-inch BENQ monitor (XL2540) with a resolution of 1920 × 1080 pixels, and a display area of 531 × 298. The refresh rate of the monitor was set to 144 Hz. The monitor was positioned 95 cm from the table edge. For a distance between the eyes and monitor of 91 cm, the monitor subtended horizontal and vertical viewing angles of 33° and 19°, respectively. The eye-tracking camera/IR light source was located at 65 cm from the head support. Participants wore closed-back headphones to block out ambient noise. There was no natural light in the room. The illuminance of the fluorescent lighting in the room near the experimental setup was around 400 lx, as measured with a Konica Minolta T-10MA illuminance meter.

Procedures

Before the experiment, participants completed a standard EyeLink nine-dot calibration procedure. Participants first looked at a number of stimuli as part of an unrelated pupillometry study lasting about 15 min (De Winter, Petermeijer, Kooijman, & Dodou, 2020). Next, the IT experiment started. Participants received task instructions on the monitor (see Figure S1 in the supplementary materials). These instructions stated that participants needed to accurately discriminate between one short and one long bar, and mentioned that the long bar would be randomly varied between the left and right positions. Furthermore, it was mentioned that participants had to press the key that matched the position of the long bar. The correct answers were the 'A' key (covered with a red sticker) if the longest leg was on the left side and the 'L' key (covered with a blue sticker) if the longest leg appeared on the right. The instructions were accompanied with an image depicting the fixation marker and an image depicting the Pi-figure with its long leg on the left, and the text "In the above example the left bar is longer, so the correct response is 'left' (key with the red sticker)". In a second instruction screen, participants were informed as follows: "This is not a reaction time task – you have as much time as you like in which to respond. You can make your response whenever you like". The experimenter provided further explanation in case the participant had questions.

Next, participants received 80 IT stimuli. The stimuli were presented in the form of videos with a frame rate of 144 frames per second.

Each video consisted of the following parts in this order (see Figure S2 for screenshots):

- A fixation marker in the middle of the screen for 500 ms
- A blank screen for 604 ms
- The Pi-figure stimulus for 14, 21, 35, 42, 56, 83, 104, or 153 ms
- A mask for 500 ms

The stimuli were drawn in MATLAB and saved as a video file having a frame rate of 144 fps. It was verified using a 1000 Hz high-speed camera that it took 2 to 3 ms for the Pi-

figure and mask to appear on the screen. Accordingly, the above exposure times of the Pi-figure were regarded as accurate. The video of the high-speed camera is available in the supplementary materials.

The legs of the Pi-figure were 124 pixels apart horizontally, which corresponds to a viewing angle of 2.2°. The short leg was 138 pixels long (2.4° vertically), and the long leg was 276 pixels long (4.8° vertically). The lines of the Pi-figure were black and 2 pixels thick. The Pi-figure was placed on a light grey background (RGB 237, 237, 237).

Although participants had been informed that they could take as much time as they wanted to respond, the maximum response time was 3.9 s (this corresponds to 5 s since the beginning of the trial minus 1.1 s, which was the elapsed time that the Pi-figure was presented). It was reasoned that this time limit would be more than sufficient for respondents to provide input.

If a participant provided a correct response, the word "CORRECT" was shown for 0.7 s, and if the participant provided an incorrect response, "INCORRECT" was shown for 0.7 s (Figure S2a). No feedback was provided if the participant did not respond.

For each of the eight exposure times, five videos showed the longer leg on the right side, and five showed the longer leg on the left side. The 80 videos were shown in a random order that was different for each participant. The experimental procedure lasted approximately 5 minutes.

After the 80 IT trials, participants answered two multiple-choice questions:

- "During the experiment I experienced (n)one of the following visual illusions:", with the following response options:

  "1 = The shorter bar of the stimulus appeared to 'grow' larger after the mask appeared."

  "2 = The mask that covered the stimulus appeared 'brighter' on the side where the bar was shortest.", and

  "3 = I have experienced no illusions at all."

- "Have you used the perceived illusion as a cue to perform the task? Choose no if no illusion was perceived)", with the response options "1 = Yes", and "2 = No".

Data Processing

Blinks were defined based on the vertical eye-gaze coordinate. Periods during which vertical eye-gaze coordinate data were unavailable, as well as periods where participants glanced above or below the edges of the screen, were labelled as 'blinks'. A manual inspection of the raw data (pupil diameter, vertical gaze coordinates) showed that the vast majority of data losses were indeed due to blinks, rather than due to looking away

from the screen. A margin of 100 ms was added before and after each blink, to account for the closing time and reopening time, respectively (Caffier, Erdmann, & Ullsperger, 2003). For each trial, data were recorded until 0.5 s after the participant provide a response. Because of the aforementioned 100 ms margin that surrounded each blink, blink data were included up to 0.4 s after the participant responded.

The following measures were calculated for each participant:

- *Non-responses*. The percentage of trials out of 80 where the participant did not respond within the allocated time.
- *Response accuracy*. The percentage of responses that were correct. Non-responses were excluded from this calculation.
- *Mean response time*. The response time was defined as the time between the onset of the Pi-figure and the moment the participant pressed one of the two keys. Non-responses were excluded from this calculation. It is noted that an IT task is not a response time task; we did not instruct participants to respond as quickly as possible. However, this does not preclude us from reporting how much time participants took to respond. We used the response time as an indicator of information processing efficiency and experience effects. Vickers, Nettelbeck, and Willson (1972) found that the mean response time decreased with exposure duration of the Pi-figure.

Note that some literature defines IT based on estimating the minimum exposure time necessary to achieve a threshold percentage of correct discriminations of the longer line (e.g., Vickers & Smith, 1986). We opted for the number of correct responses as a simpler and more tractable performance score (e.g., Posthuma, De Geus, & Boomsma, 2001; Ritchie et al., 2013). Furthermore, it was impossible to calculate a threshold percentage because some participants showed poor performance (e.g., for 27 of 147 participants, less than 60% of responses were correct).

A preliminary analysis of the horizontal and vertical eye gaze coordinates revealed no noteworthy patterns between trials in which participants provided a correct response and trials in which the participants provided an incorrect response. In short, it was found that participants, on average, looked about 15 pixels more downward at the moment of stimulus presentation for trials with an incorrect response as compared to trials with a correct response. We suspect that this small effect is confounded with partial eye closures, causing an apparent downward movement of the vertical gaze coordinate. Because eye-movement effects appeared to be small and not of general interest, they were not pursued further.

We calculated associations between the performance measures and trial number, self-reported illusion, and percentage of trials in which the participant was blinking (for

distinct elapsed times during the trial: 0, 0.22, 0.44, 0.66, 0.88, and 1.10 s). Group comparisons for the illusions were performed using unequal-variances *t*-tests (Welch's tests). Cohen's *d* was used as an effect size measure. Associations between the response accuracy and the percentage of trials in which the participant blinked were computed at the level of participants, using two complementary measures: Pearson's product-moment correlation coefficient (*r*) and Spearman's rank-order correlation coefficient (ρ). Pearson's correlation is a measure of the degree of linear association. It is intuitively interpretable but has the disadvantage of being less stable when outliers are present or when the distribution is heavy-tailed. Spearman's correlation, on the other hand, is robust to outliers and tailed distributions (De Winter, Gosling, & Potter, 2016).

Follow-up Experiment

The above experiment had a number of characteristics that may have made the task difficult or confusing for participants. A follow-up experiment was conducted, with the goal to examine whether the results replicated in improved experimental conditions. The follow-up experiment was the same as the above experiment, but with the following modifications:

- **New participants.** Participants were 165 MSc engineering students. Six participants had to be removed due to missing or low-quality eye-tracking data. The remaining 159 participants were 50 females and 109 males with a mean age of 23.52 years (*SD* = 1.98). Thirty-three participants used visual aids during the experiment (23 contact lenses, 10 glasses).
- **Enhanced instructions.** The instructions were expanded. More specifically, it was mentioned: "In each trial, you will be shown (one after the other): 1) a fixation marker, 2) then, a stimulus consisting of two lines of different lengths, 3) shortly after, the lines will be masked". The fixation marker, stimulus (Pi-figure), and mask were each accompanied with a figure. On a second instruction screen, a Pi-figure with its long leg on the left, and a Pi-figure with its long leg on the right were shown. The accompanying text stated: "Your task is to indicate which of the two lines of the stimulus was the longest. If the left line was longer, press the left shift key. If the right line was longer, press the right shift key".
- **Practice trials.** The first experiment did not contain practice trials. In the follow-up experiment, we included three practice trials, with exposure times of 1000 ms, 500 ms, and 201 ms. After each practice trial, the correct answer was shown on the screen.
- **Uniform luminance.** In the first experiment, the feedback screens stating "CORRECT" or "INCORRECT" were brighter than the stimulus trials (see Figure S2a). This increased brightness may have given rise to exogenous eye blinks. In the follow-

up experiment, we used a uniform grey background (RGB 127, 127, 127) for the entire IT task, including the feedback screens. Furthermore, the "CORRECT" and "INCORRECT" messages were shown in an outline font to minimize the overall effect of luminance (Figure S2b).

- **Adjustments to the stimulus.** The blank screen before the presentation of the Pi-figure was shown for 500 ms instead of 604 ms to reduce participants' waiting time. Furthermore, the lines of the Pi-figure were made thicker (6 pixels) to increase the likelihood that participants could see the Pi-figure. The legs of the Pi-figure were 90 pixels apart horizontally, which corresponds to a viewing angle of 1.6°. The short leg was 90 pixels long (1.6° vertically), and the long leg was 180 pixels long (3.1° vertically).
- **No response time limit.** In the first experiment, participants had to respond within a time limit of 3.9 s. In the follow-up experiment, participants had infinite time to respond. Because of this, non-responses were impossible, and we expected that this would reduce confusion among participants. Because long response times (i.e., outliers) were possible, the median response time, instead of the mean response time, was used as a measure.
- **Enhanced questions about illusions.** In the first experiment, we inquired whether participants perceived an apparent movement illusion or a brightness illusion. In the follow-up experiment, we asked about visual illusions in greater detail, based on Alexander and Mackenzie (1992). More specifically, we asked whether participants experienced (1) the ends of the lines of the stimulus moving/stretching downward upon the appearance of the mask, (2) a flash arising from the stimulus upon the appearance of the mask, (3) a small black gap at the ends of the stimulus lines upon the appearance of the mask, (4) another type of visual illusion, or (5) no illusion. Additionally, participants were required to answer "what exactly did you perceive, and did you use the illusions to perform better at the task?" using a textbox field.
- **Inclusion of cognitive test.** The first experiment did not include an IQ test. To examine the criterion validity of the IT task, we added a 12-item version of Raven's advanced progressive matrices (Arthur, Tubre, Paul, & Sanchez-Ku, 1999). Participants had 7 minutes to solve as many of the 12 items as they could. The Raven's matrices were completed after the IT task, and participants entered their responses using the keyboard.
- **Binocular eye-tracking.** We used a newer version of the EyeLink 1000 Plus eye-tracker, which measured eye-movements of both eyes at 2000 Hz. A blink was defined as in the first experiment, except that the gaze data for the two eyes were

first averaged. If vertical gaze data for one of the two eyes was unavailable, then this was counted as a blink.

RESULTS

The 147 participants each performed 80 IT trials. On average, participants had 4.74 non-responses (*SD* = 9.88). Accordingly, the average number of responses per participant was 75.26.

Experience Curves

Figure 1 shows the percentage of 80 trials where a response was provided within the allocated time, the response accuracy (i.e., the percentage of responses that were correct), and the mean response time. With increasing experience, response accuracy increased, response time decreased, and the likelihood of giving a response within the available time increased. The latter measure reached an early plateau at about ten trials. The mean response time, however, kept reducing with trial number (Figure 1). There were no significant associations between participants' response accuracy and age (r = –0.07, p = 0.399) or gender (r_{pb} = 0.14, p = 0.094, coded as 1 = male, 2 = female).

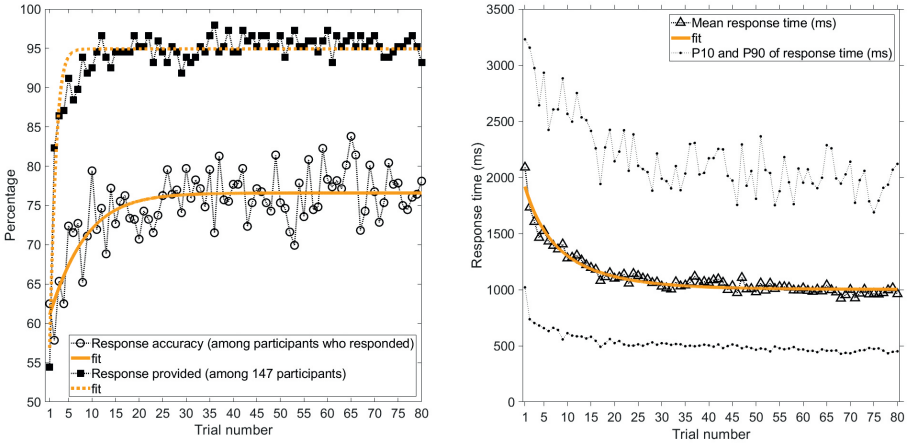


Figure 1. Experience curves as a function of trial number, where Trial 1 is the first IT stimulus presented, and Trial 80 is the last IT stimulus presented. Left = Response accuracy (i.e., percentage of responding participants who provided a correct response), and percentage of 147 participants who provided a response within the allocated time. Right = Mean response time of the participants who provided a response, together with 10th and 90th percentiles. Exponential fits, $y = 1/(a + b*\exp(-c*x))$, are shown where x is the trial number, and a, b, and c are fitted parameters. For the 'response accuracy' curve, a = 0.0131, b = 0.00387, c = 0.155 (r^2 = 0.53). For the 'response provided' curve, a = 0.0105, b = 0.0198, c = 1.033 (r^2 = 0.87). For the 'mean response time' curve, a = 0.00100, b = –0.000516, c = 0.0811 (r^2 = 0.94). Note that the IT stimuli were presented in random order.

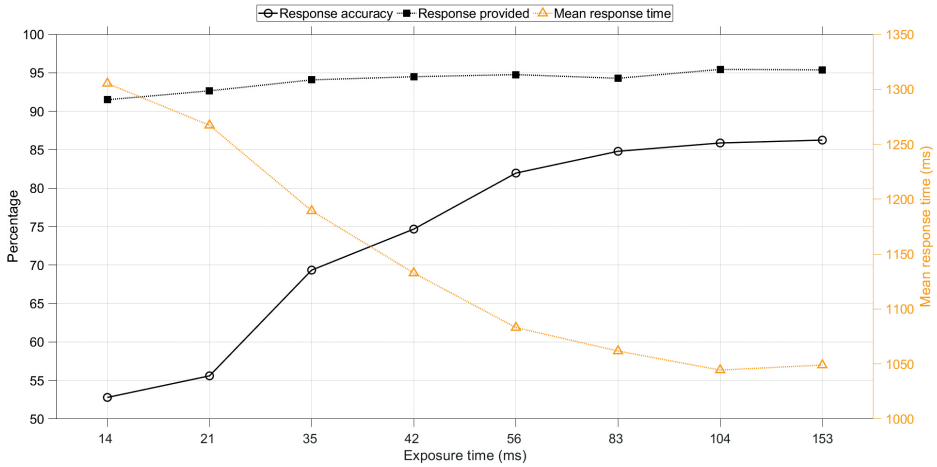


Figure 2. Response accuracy (i.e., percentage of responses that were correct), percentage of trials with a response within the allocated time, and mean response time as a function of exposure time of the Pi-figure. The means and standard deviations are provided in Table S1.

Effect of Exposure Time

Longer exposure times yielded a higher response accuracy and a faster response (Figure 2). More specifically, when the exposure time was low (14 ms), response accuracy was barely above chance level, and when the exposure time was high (153 ms), the response accuracy was 86.2%. The mean response time decreased from 1305 ms for a 14 ms exposure time to 1049 ms for a 153 ms exposure time. Participants also were more likely to respond within the available time limit when the exposure time was higher.

Association with Self-Reported Illusions

Next, we examined the association between IT performance and self-reported illusions. The brightness illusion was relatively infrequent (17 participants, 12%) as compared to the growing illusion (56 participants, 38%) and no illusion (74 participants, 50%). Of the 17 participants who experienced the brightness illusion, 14 (82%) reported using this illusion as a cue to perform the task. Of the 56 participants who experienced the growing illusion, 47 (84%) reported using this illusion as a cue to perform the task.

Results in Table 1 show that the brightness illusion is associated with a higher response accuracy, a lower percentage of non-responses, and a faster mean response time as compared to no illusion. The effects are illustrated using boxplots in the supplementary materials (Figures S3a, S4a, S5).

Attention during the Trials

In total, 11760 trials were completed (147 participants × 80 trials per participant). A keypress response, either correct or incorrect, was recorded in 11063 of those trials. Blinking data for 19 of those 11063 trials were excluded because there were not

enough ocular data. More specifically, participants in those 19 trials were observed to be blinking for over 50% of the time of that trial, which could be explained because of poor eye tracking quality or participants not looking at the screen.

Figure 3 shows the percentage of trials with blinking as a function of elapsed time during the IT trial. A distinction is made between 8308 trials in which a participant provided a correct response and 2736 trials in which the participant provided an incorrect response.

Two main patterns can be distinguished. First, the blinking patterns were highly dynamic. Participants hardly blinked during the crucial period of the presentation of the Pi-figure, and they blinked after the trial had ended (Figure 3, top). Second, there is a distinction between blinking patterns of correct and incorrect responses. Incorrect responses were associated with blinking when the Pi-figure was presented, whereas correct responses were associated with blinking afterwards (Figure 3, top).

The high blink rates for correct responses after the presentation of the Pi-figure can be explained by the fact that participants responded about half a second faster for correct responses compared to incorrect responses (M = 982 ms for the 8322 correct responses, M = 1461 ms for the 2741 incorrect responses). As can be seen in Figure 3, bottom, many participants blinked after having responded.

Table 1. Inspection time task performance per self-reported illusion (N = 147).

	Response accuracy (%) of trials in which the participant responded)	No response (% of all 80 trials)	Mean response time (ms)
	Mean (*SD*)	**Mean (*SD*)**	**Mean (*SD*)**
Growing illusion (n = 56)	74.52 (14.23)	3.98 (5.74)	1141 (522)
Brightness illusion (n = 17)	80.33 (7.76)	0.82 (1.01)	860 (273)
No illusion (n = 74)	72.33 (16.87)	6.22 (12.81)	1202 (560)
	Welch's test	**Welch's test**	**Welch's test**
Growing vs. no illusion	$t(126.4) = 0.80$, $p = 0.426$	$t(106.8) = 1.33$, $p = 0.185$	$t(122.4) = 0.64$, $p = 0.523$
Brightness vs. no illusion	$t(55.3) = 2.94$, $p = 0.005$	$t(76.8) = 3.57$, $p < 0.001$	$t(51.3) = 3.68$, $p < 0.001$

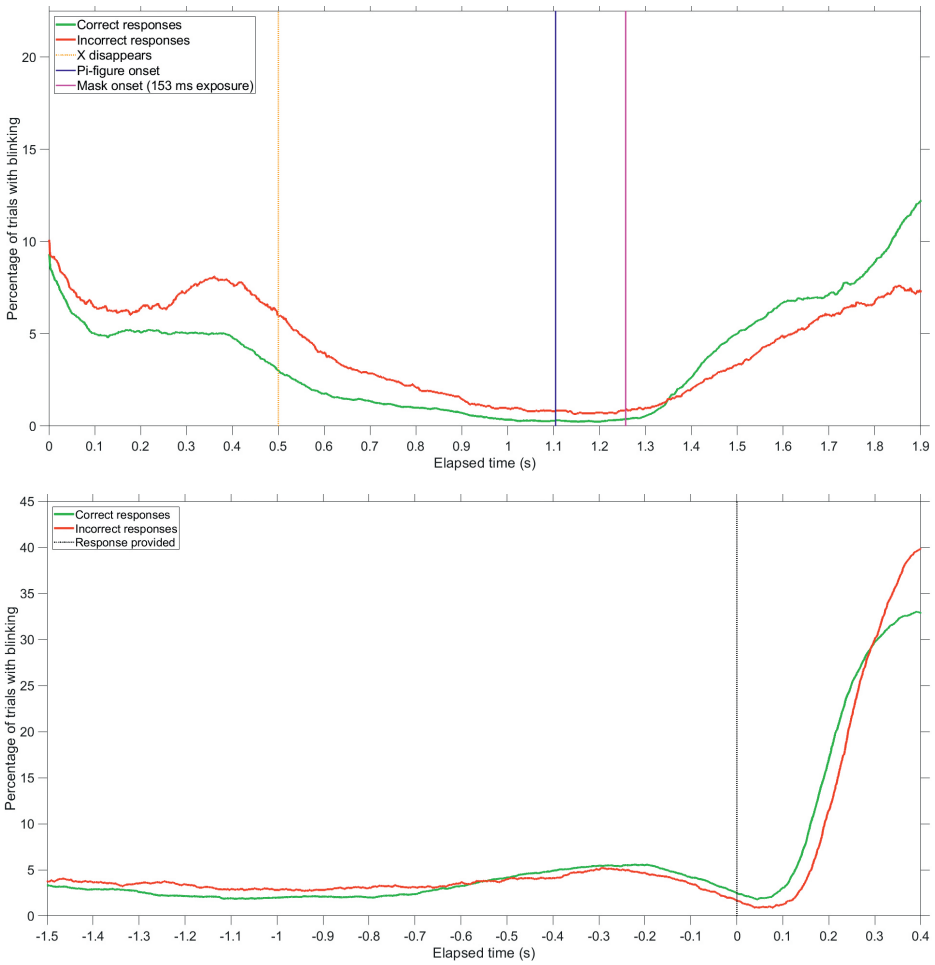


Figure 3. Percentage of trials in which participants were blinking, for each time sample. A distinction is made between trials where participants provided a correct response (n = 8308) and trials where participants provided an incorrect response (n = 2736). Top figure: results time-locked to the stimulus (occurring at t = 1.1 s). Vertical lines are shown for the moment the fixation marker (X) disappeared, the moment the Pi-figure was presented, and the moment the mask was presented for the maximum exposure time of 153 ms. Bottom figure: results time-locked to the participants' response, indicated by the vertical line at t = 0 s. Participants were provided with a "CORRECT" or "INCORRECT" feedback message after responding. Data were included up to 0.4 s after the participant provided a response; therefore, the number of data points near the end of the top figure (t = 1.9 s) or the beginning of the bottom figure (t = –1.5 s) is reduced (n = 8163 for correct responses, n = 2705 for incorrect responses).

The individual differences in blinking are illustrated using scatter plots at the level of participants, see Figure 4. It can be seen that a negative correlation exists between blinking and the percentage of correct responses.

Blinking as a Function of Trial Number

Figure 5 shows the percentage of blinking as a function of elapsed time during the IT trial. A distinction is made between the degree of task experience, by creating 8 groups of 10 trials. It can be seen that during the first ten trials, participants relatively often blinked during the presentation of the Pi-figure. At later trials, participants blinked more and more after the stimulus presentation.

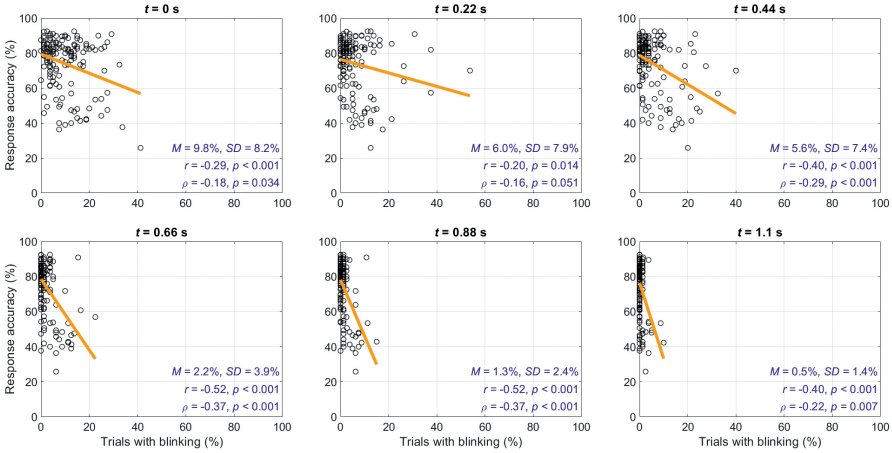


Figure 4. Response accuracy (i.e., percentage of responses that were correct) versus the percentage of trials with blinking at the level of participants (N = 147), per 220 ms of elapsed time into the trial. Also shown is a least-squares regression line, means and standard deviations of the percentage of trials with blinking, and the Pearson correlation coefficients (r) and Spearman rank-order correlation coefficients (ρ). The fixation marker onset occurs at t = 0 s. The onset of the Pi-figure occurs at t = 1.1 s.

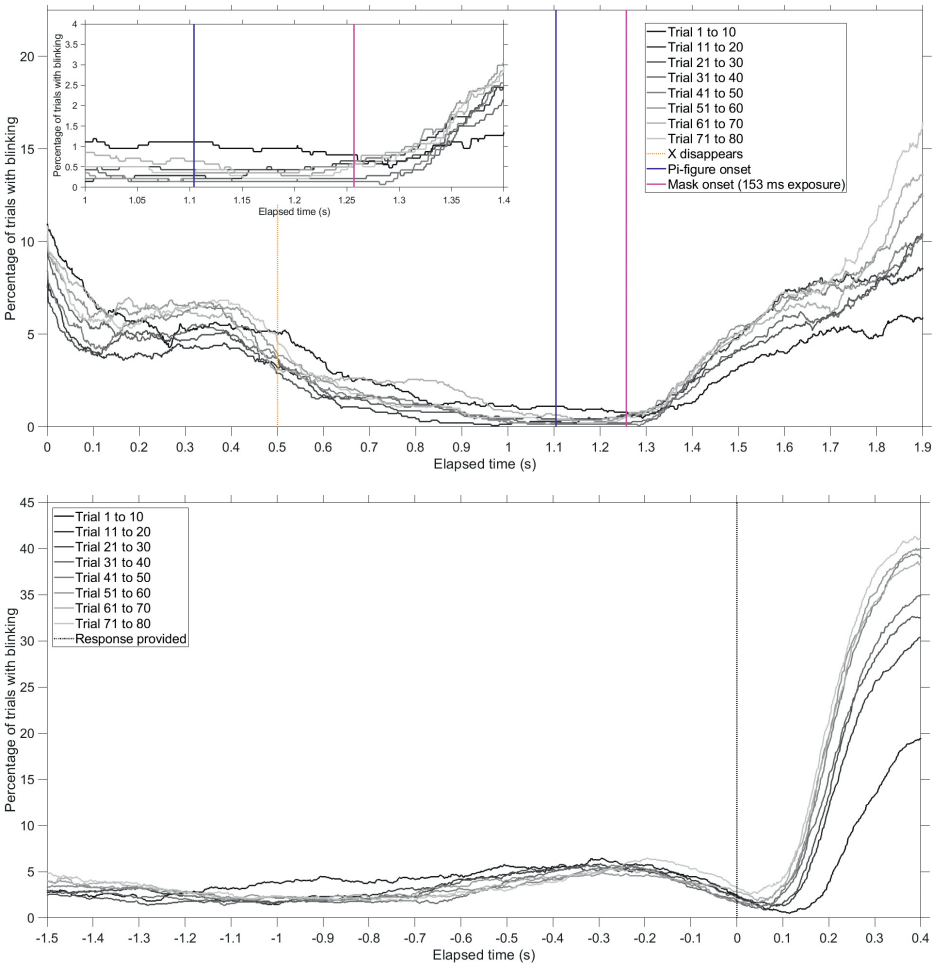


Figure 5. Mean percentage of all trials in which the participant was blinking, per group of 10 trials (11044 trials in total). Top: results time-locked to the stimulus (occurring at t = 1.1 s). An inset is provided for elapsed times between 1.0 s and 1.4 s. Vertical lines are shown for the moment the fixation marker (X) disappeared, the moment the Pi-figure was presented, and the moment the mask was presented for the maximum exposure time of 153 ms. Bottom: results time-locked to the participants' response, indicated by the vertical line at t = 0 s. Participants were provided with a "CORRECT" or "INCORRECT" feedback message after responding. Data were included up to 0.4 s after the participant provided a response.

Follow-up Experiment

In the follow-up experiment, learning curves can be seen, similar to the learning curves of the first experiment, see Figure 6. For the response time, the fit was still strong (r^2 = 0.95). The response accuracy, however, showed a less strong learning curve as compared to the first experiment. The response accuracy was considerably higher as compared to the first experiment, with a score of 98.6% for the highest exposure duration, compared to 86.2% in the first experiment (Figure 7). Participants also

responded substantially faster, with a median response time for the highest exposure duration of 562 ms versus 958 ms in the first experiment (see Tables S1 and S2). There were no significant associations between participants' response accuracy and age (r = 0.00, p = 0.960), and females had a slightly lower response accuracy than men (r_{pb} = –0.18, p = 0.026, coded as 1 = male, 2 = female).

Exponential fits, $y = 1/(a + b*\exp(-c*x))$, are shown, where x is the trial number, and a, b, and c are fitted parameters. For the 'response accuracy' curve, a = 0.0114, b = 0.00158, c = 0.0895 (r^2 = 0.40). For the 'median response time' curve, a = 0.00188, b = –0.00143, c = 0.1433 (r^2 = 0.95). Note that the IT stimuli were presented in random order.

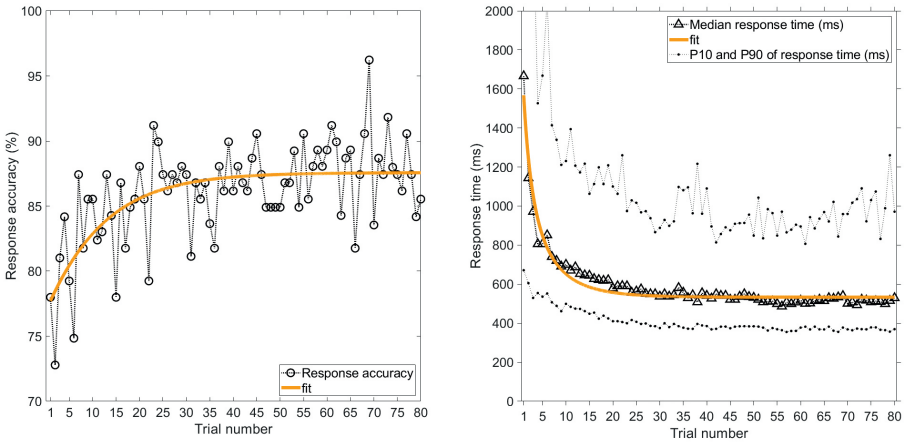


Figure 6. Follow-up experiment: Experience curves as a function of trial number, where Trial 1 is the first IT stimulus presented, and Trial 80 is the last IT stimulus presented. Left = Response accuracy (i.e., percentage of 159 participants who provided a correct response). Right = Median response time among 159 participants, together with 10th and 90th percentiles.

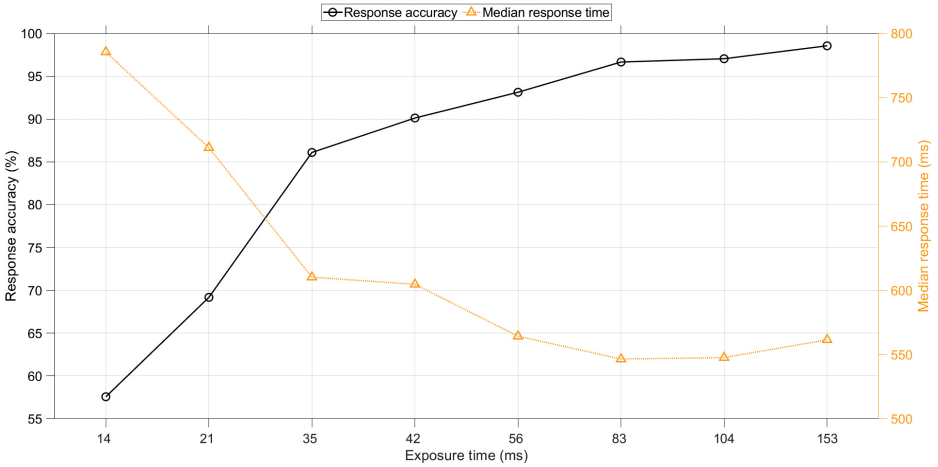


Figure 7. Follow-up experiment: Response accuracy and median response time as a function of exposure time of the Pi-figure. The median response time was calculated per participant per exposure time and subsequently averaged over the 159 participants. The means and standard deviations are provided in Table S2.

Table 2. Follow-up experiment: Inspection time task performance per self-reported illusion (N = 159).

	Response accuracy (% of trials)	**Median response time (ms)**
	Mean (*SD*)	**Mean (*SD*)**
Moving/stretching illusion (n = 49)	85.65 (7.76)	629 (176)
Flash illusion (n = 39)	84.70 (8.50)	609 (217)
Black gap illusion (n = 8)	87.81 (4.47)	549 (81)
Other illusion (n = 25)	87.04 (9.40)	530 (104)
No illusion (n = 38)	86.90 (6.15)	552 (129)
	Welch's test	**Welch's test**
Moving/stretching vs. no illusion	$t(85.0) = 0.83$, $p = 0.406$	$t(84.7) = 2.33$, $p = 0.022$
Flash vs. no illusion	$t(69.3) = 1.30$, $p = 0.196$	$t(62.2) = 1.40$, $p = 0.167$
Black gap vs. no illusion	$t(13.3) = 0.49$, $p = 0.632$	$t(15.5) = 0.09$, $p = 0.932$
Other vs. no illusion	$t(37.5) = 0.07$, $p = 0.946$	$t(58.3) = 0.75$, $p = 0.457$

In the follow-up experiment, the response accuracy and the response times were similar between participants who reported having perceived an illusion and participants who reported having perceived no illusion (see Table 2). In other words, the association between illusions and task performance, as observed in the first experiment (see Table 1), did not replicate. In fact, the results showed that participants who perceived the moving/stretching illusion had a significantly *longer* response time than those who perceived no illusion.

An examination of the responses to the free-response item showed that 157 of 159 participants provided a meaningful response. The responses varied considerably, with many participants reporting no illusion, or describing general phenomena ("only experienced an effect similar to tunnel vision") rather than illusions related to task performance. However, several interesting observations were made:

- 27 participants reported that a change occurred on the shorter side of the Pi-figure in particular, e.g. "I looked at where the moving appeared and then I knew that the other side would have the longer end".
- 9 participants reported that the shorter line moved more slowly than the longer line, e.g. "The lines seemed to stretch downwards. The line which stretched down faster was the longer line, so yes I used it to perform the task".
- 7 participants reported aftereffects, e.g. "I could see the lines on the screen even after the image was gone and it helped me predict some answers correctly". However, in some cases the aftereffects were described as having a negative effect on performance, e.g. "I did, however, experience an after-image which made it difficult to identify this contrast, the more tired my eyes became".
- 4 participants reported that a flash occurred on one side, e.g. "I saw a flash on the shorter side when the mask appears". For three participants, the flash occurred on the shorter side; for one participant it occurred on the longer side.

It is of note that several participants reported relying on the illusion ("perceived moving stimulus, kind of amazed that I sometimes did not really see the whole stimulus but knew what side it was, left or right") while others stated that they saw no illusion whatsoever ("I did not see any illusions, I am just really good at this"). Interesting as well, some participants reporting seeing no illusion in the multiple-choice item, but still referred to a change or motion, e.g., "stretching of line on the shorter side; therefore, the other side should have been the longest line" or "the longest line didn't move as much … so, it was the side with little movement".

In total, 12720 trials were completed (159 participants × 80 trials per participant), with response data and eye-tracking being available for 12683 trials. Eye-blinking patterns

showed a similar pattern as in the first experiment, with participants avoiding blinking at the moment of the presentation of the Pi-figure, and blinking after that (Figure 8). Again, correct responses were associated with not blinking during the moment of Pi-figure presentation.

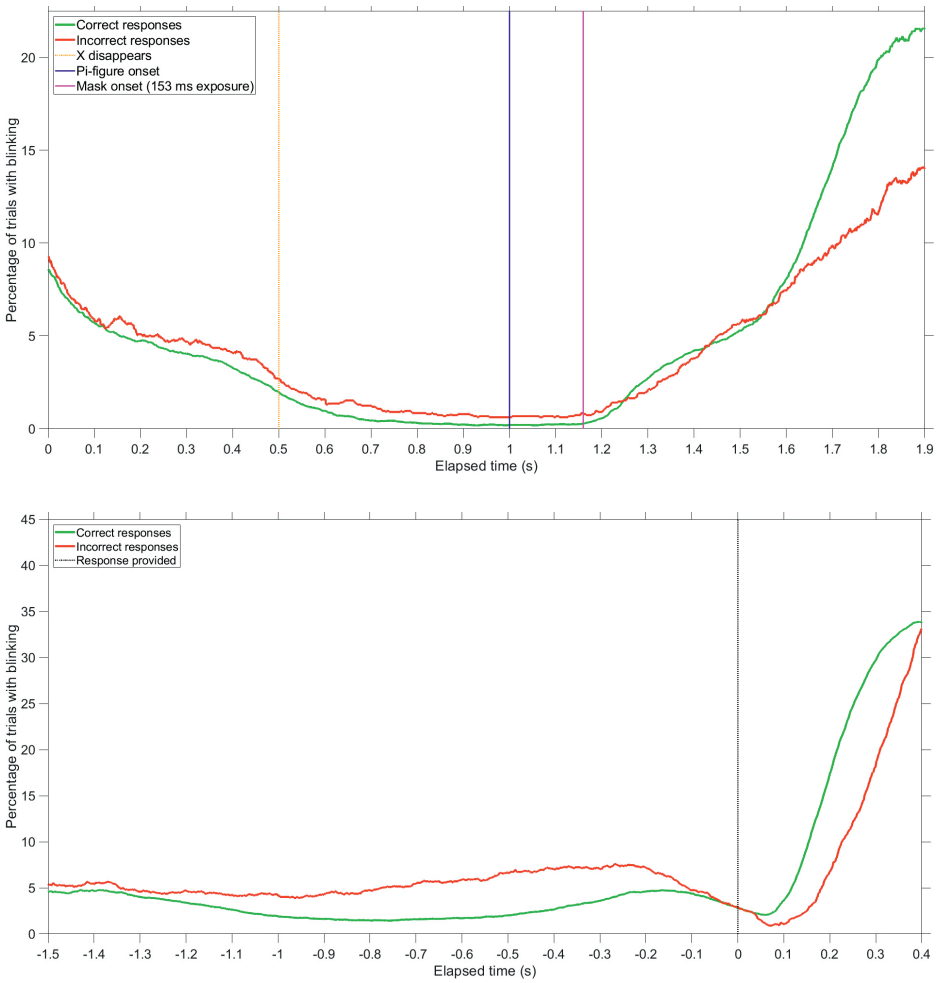


Figure 8. Follow-up experiment: Percentage of trials in which participants were blinking, for each time sample. A distinction is made between trials where participants provided a correct response (*n* = 10915) and trials where participants provided an incorrect response (*n* = 1768). Top figure: results time-locked to the stimulus (occurring at *t* = 1.0 s). Vertical lines are shown for the moment the fixation marker (X) disappeared, the moment the Pi-figure was presented, and the moment the mask was presented for the maximum exposure time of 153 ms. Bottom figure: results time-locked to the participants' response,indicated by the vertical line at t = 0 s. Participants were provided with a "CORRECT" or "INCORRECT" feedback message after responding. Data were included up to 0.4 s after the participant provided a response; therefore, the number of data points near the end of the top figure (*t* = 1.9 s) or the beginning of the bottom figure (*t* = –1.5 s) is reduced (*n* = 7967 for correct responses, *n* = 1569 for incorrect responses).

However, the associations, as shown in Figure 9, were weaker as compared to the first experiment. It can also be seen from Figure 9 that there were only few trials in which participants blinked when the Pi-figure was shown (0.3% at 0.88 s and 0.3% at 0.3% at 1.1 s) in comparison to the first experiment (1.3% at 0.88 s and 0.5% at 1.1 s).

Finally, it was found that the overall response accuracy on the IT task (M = 86.04%, SD = 7.73%) correlated significantly (r = 0.18, p = 0.027) with the number of items that participants got correct on the Raven matrices (M = 7.30, SD = 1.93). This finding demonstrates some validity of the IT task as a predictor of performance on the Raven matrices task.

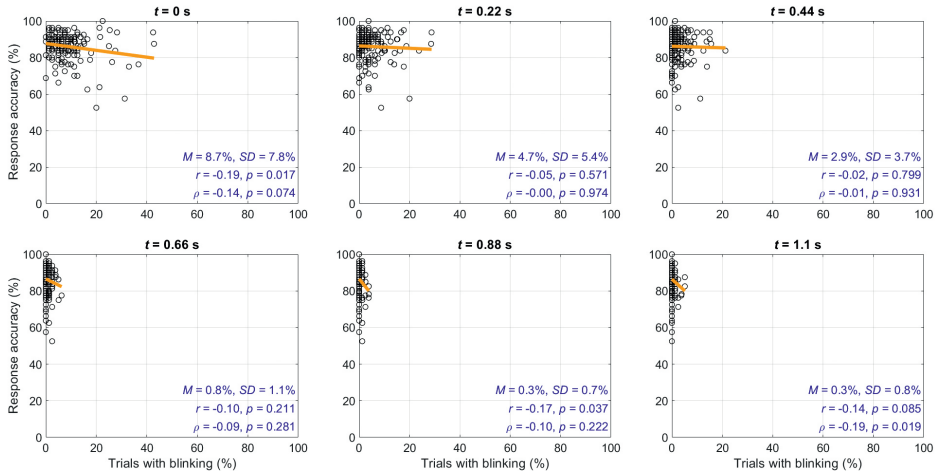


Figure 9. Follow-up experiment: Response accuracy (i.e., percentage of trials with a correct response) versus the percentage of trials with blinking at the level of participants (N = 159), per 220 ms of elapsed time into the trial. Also shown is a least-squares regression line, means and standard deviations of the percentage of trials with blinking, and the Pearson correlation coefficients (r) and Spearman rank-order correlation coefficients (ρ). The fixation marker onset occurs at t = 0 s. The onset of the Pi-figure occurs at t = 1.0 s.

DISCUSSION

Experience Effects and Overall Performance

The results of the first experiment showed that IT performance improved with trial number. In the follow-up experiment with improved task instructions and the inclusion practice trials, learning curves were still present. The shapes of the experience curves suggest that participants, in the aggregate, required about ten trials to get familiar with the task, after which they increased their attention to the task and reduced their response latency. The observed experience curves match previous research showing that the IT of children improves across sessions and testing days (Nettelbeck & Vita, 1992). Similarly, Bors et al. (1999) and Blotenberg and Schmidt-Atzert (2019) found that

participants performed better when completing the IT session for a second or third time as compared to the first time.

Participants were aided with knowledge-of-results feedback, which can be expected to have contributed to improved performance as compared to not receiving such feedback (Salmoni, Schmidt, & Walter, 1984). Also, our study was conducted with MSc students at an engineering university, who are expected to have above-average IQs, presumably in the 115–130 range (based on Wai, Lubinski, & Benbow, 2009).

Despite the task feedback and presumably high intelligence of participants, performance in the first experiment was low, with a response accuracy of 86.2% for the highest exposure time. In addition, there were a considerable number of non-responses, especially in the first few trials. The low accuracy as well as non-responses can be explained by the fact that we provided participants with only basic instructions, no practice trials, and no performance feedback if the participants did not respond. In the follow-up experiment which included enhanced instructions, practice trials, and no response-time limit, a near-perfect response accuracy of 98.6% was obtained for the highest exposure time.

Our results point to the importance of making sure that participants understand the task. Previous IT studies have been conducted with different population groups, including children (e.g., Anderson, 1986; Nettelbeck & Young, 1990) and old persons (Johnson & Deary, 2011), which makes us wonder whether participants in all cases have understood the task. It seems plausible that the link between IT and IQ can, in part, be explained by the fact that persons with higher IQ are more likely to understand what they have to do while performing the IT task.

Visual Illusions

The first experiment showed that IT performance is better among participants who reported a brightness illusion than among those who reported no illusion. These findings confirm previous research (e.g., Mackenzie & Bingham, 1985) regarding the benefit of perceiving illusions, with the difference that our study showed that the brightness illusion yielded a statistically significant benefit. In contrast, previous research was mostly concerned with the apparent movement illusion (see Introduction).

About 83% of participants who reported a visual illusion indicated using this illusion as a cue to perform the task. It is possible that participants intelligently deployed this cue for selecting the response key that was on the opposite side of the illusion. Egan (1994) explained: “Once the subject has become aware of this motion, s/he need only register the aftereffect, then press the response key on the side opposite to the region of motion.” (p. 307). The self-reports in the follow-up experiment indicated that

participants did use such intelligent strategies, although the content of the responses varied considerably.

As pointed out above, visual illusions may cause one to employ a strategy that increases performance. However, our results suggest two additional explanations for the perception of illusions. First, the self-reports of the follow-up experiment indicate that what to consider an illusion is to some extent a matter of semantics. Some participants recognized the change from Pi-figure to the mask as a stretching/movement illusion, whereas other participants appeared to describe the same stretching/movements and did not regard it as an illusion, but merely as a change from one image to the other. Our observations appear to be in line with Simpson and Deary (1997) who found no causal effect of 'macrolevel' strategy use on IT and concluded that strategies are a verbalization of 'microlevel' cognitive processes.

A second explanation for the perception of illusions is that they are a by-product of understanding the task and knowledge of where to look. Conversely, if one does not understand the task or if one fails to distinguish the legs of the Pi-figure, then no illusion is likely to be perceived. An explanation for the superior performance of strategy users as an epiphenomenon has been considered before. Egan and Deary (1992), for example, argued that perceived illusions are "simply something seen when a discrimination is still possible for a subject at a short absolute IT duration" (p. 164). The standard deviations of the response accuracy, non-response percentage, and the mean response times were considerably smaller for the brightness illusion group as compared to the other two groups, which can be explained by the fact that a number of participants performed very poorly, sometimes around chance level (Figure S3a) or did not respond at all (Figure S5). These poor performers may have misunderstood the task or may have failed to see the legs of the Pi-figure. Spontaneous remarks by the participants reinforce the idea that the IT task was regarded as confusing. For example, a number of participants indicated that they thought they had to detect the difference in the lengths of the legs of the mask (while apparently not having seen the Pi-figure at all). In the follow-up experiment, we found no significant differences in task performance between four categories of strategy use versus no reported strategy use, and no incidences of extremely poor performance (Figure S3b). This finding reinforces the epiphenomenal explanation. In summary, the reporting of the brightness illusion may be a by-product of understanding the task or concentration at the task. It may even be hypothesized that the apparent motion illusion is a completely normal phenomenon that can be experienced by everyone, similar to the illusion of motion that occurs when playing the pictures of a movie at a minimal frame rate (Holcombe, 2009).

Alexander and Mackenzie (1992) reported four possible illusions: apparent motion, flash-brightness, ends-stand-out, and after-image, whereas Egan (1994) reported movement,

flickering, and brightness. In our follow-up experiment, participants revealed interesting refinements to these illusions, with some referring, for example, to the fact that the short leg of the Pi-figure moved slower than the longer leg. Multiple-choice questions and free-response items, as used in the present study, provide only limited information about strategy use. For future research, we recommend performing interviews to examine how participants perceived and used the illusions. This recommendation is in line with Egan and Deary (1993), who advised "continuous monitoring of self-reports to describe the 'on-line' natural history of strategy development" (p. 135).

Attention

Using eye-tracking equipment, we found that the IT task is highly dynamic: participants avoided blinking at the critical moment of the presentation of the Pi-figure. The overall increase in blinking with trial number, as shown in Figure 5, may have been caused by fatigue or eyestrain. In the first experiment, the correct/ incorrect feedback (see Figure S2a) was bright and resulted in reflexive pupil constriction (see Figure S6), and may have contributed to a reflexive blinking response. However, in the follow-up experiment, with tight luminance control, many participants also blinked after the presentation of the Pi-figure, suggesting that this blinking is due to post-trial relaxation rather than due to a light reflex. In summary, participants in the first experiment and the follow-up experiment made sure that they were hardly blinking during the presentation of the Pi-figure, pointing to a crucial role of visual attention management while performing the IT task.

We found that whether one blinks at a particular moment of the trial was related to response accuracy. The corresponding correlations were stronger in the first experiment (ρ between –0.25 and –0.40) than in the follow-up experiment (ρ between –0.10 and –0.20). This difference can be explained by the larger individual differences in blinking and response accuracy in the first experiment, where some participants performed very poorly and blinked in a substantial number of trials when the Pi-figure was shown. Of note, the correlations are almost as strong as the correlation between IQ and IT, which Grudnik and Kranzler (2001) using meta-analysis estimated at –0.30 (uncorrected for range restriction and measurement error). Our IT-blinking correlations confirm early small-subject research of Nettelbeck et al. (1986), who found that a low-ability participant group (low-IQ participants, who obtained long IT scores) exhibited more blinking than a control group.

How should the correlation between blinking and IT be interpreted? On the one hand, it may be regarded as self-evident that blinking correlates with IT because if no light falls on the retina, better than chance performance is physically impossible. However, the blocking of light cannot be the only explanation of the observed IT-blinking correlations because only in a small number of trials (<1%) did the participants blink during stimulus

presentation. Hence, blinks are not just a direct cause of poor IT performance, but also indicative of attention during the experiment in general. This is consistent with the above-mentioned epiphenomenal explanation of perceiving visual illusions: if not understanding the task or not knowing when/where the look, then blinking may be expected at inappropriate moments and performance may be expected to be poor.

Our work showed that IT is associated with motor activity of the eyelids, where motor activity refers to blinking after the presentation of the Pi-figure and blink inhibition when the Pi-figure is visible. The involvement of motor activity would be in contradiction to, amongst others, Jensen (2006), who stated that IT is captured "independently of the whole efferent aspect of RT" (p. 84). Not only blinking but also inhibition of blinking involves certain mental demands. An fMRI study by Chung Yoon, Song, and Park (2006) showed that voluntary and inhibited eye blinks involve the precentral gyrus, a region of the brain concerned with the coordination of movement. Berman Horovitz, Morel, and Hallett (2012) found, also using fMRI, that suppression of blinks is associated with a wide network of brain activations associated with the build-up of bodily urge.

Conclusions and Recommendations

Our research contributes to the view that there is a multitude of factors associated with such a simple task as IT, including focused attention, the perception of illusions, understanding of the task, and task experience. These findings reject the hypothesis that IT is a univariate construct, and suggest that previously documented IT-IQ correlations are because of multiple overlapping processes (Kovacs & Conway, 2016; Spearman, 1923) rather than pure mental speed (see also Stankov, 2004).

A limitation of our study is that each participant completed only 80 IT trials and that long-term learning was not assessed. Another limitation is that our sample consisted of university students only. Although the use of university students appears to be common in IT research (Deary, Caryl, Egan, & Wight, 1989; Grudnik & Kranzler, 2001), a more heterogeneous sample can be expected to cause disattenuated correlations between IT and attention. Finally, it would be interesting to examine whether our findings regarding attention generalize to other types of elementary cognitive tasks. Johns et al. (2009) previously reported associations between blinking and visual reaction times. We expect that visual attention can explain a portion of the variance in task performance in psychometric tests.

In our follow-up experiment, we observed a modest correlation of 0.18 between IT and performance measured using a short version of Raven's advanced progressive matrices. This correlation may become stronger if using a more heterogeneous pool of participants. Also, we recommend that future experiments include more participants and a full IQ test. It would be worthwhile to examine how task experience and blinking are associated with intelligence.

Finally, it would be useful to examine what display characteristics contribute to performance and criterion validity. Early studies used bright LED displays (Egan, 1994), whereas we used a grey background on a computer monitor. The use of computer screens has been criticized (Simpson & Deary, 1997), but display technologies have developed significantly over the last decades, now offering high refresh rates. It is possible that low contrast displays emphasize the factors the psychometrician is interested in, such as sensory speed, perceptual coding, or attentional processes (Levy, 1992). On the other hand, perhaps low contrast displays dilute the measurement of the speed of information intake as determined by, for example, nerve conduction velocity (Miller, 1994).

REFERENCES

Adam, K. C. S., & deBettencourt, M. T. (2019). Fluctuations of attention and working memory. *Journal of Cognition*, *2*, 33. DOI: https://doi.org/10.5334/joc.70

Alexander, J. R. M., & Mackenzie, B. D. (1992). Variations of the 2-line inspection time stimulus. *Personality and Individual Differences*, *13*, 1201–1211. DOI: https://doi.org/10.1016/0191-8869(92)90256-O

Anderson, M. (1986). Inspection time and IQ in young children. *Personality and Individual Differences*, *7*, 677–686. DOI: https://doi.org/10.1016/0191-8869(86)90037-1

Anderson, M. (1989). The effect of attention on developmental differences in inspection time. *Personality and Individual Differences*, *10*, 559–563. DOI: https://doi.org/10.1016/0191-8869(89)90038-X

Anderson, M., Reid, C., & Nelson, J. (2001). Developmental changes in inspection time: What a difference a year makes. *Intelligence*, *29*, 475–486. DOI: https://doi.org/10.1016/S0160-2896(01)00073-3

Arthur, W., Jr., Tubre, T. C., Paul, D. S., & Sanchez-Ku, M. L. (1999). College-sample psychometric and normative data on a short form of the Raven Advanced Progressive Matrices Test. *Journal of Psychoeduca-tional Assessment*, *17*, 354–361. DOI: https://doi.org/10.1177/073428299901700405

Berman, B. D., Horovitz, S. G., Morel, B., & Hallett, M. (2012). Neural correlates of blink suppression and the buildup of a natural bodily urge. *NeuroImage*, *59*, 1441–1450. DOI: https://doi.org/10.1016/j.neuroimage.2011.08.050

Blotenberg, I., & Schmidt-Atzert, L. (2019). On the locus of the practice effect in sustained attention tests. *Journal of Intelligence*, *7*, 12. DOI: https://doi.org/10.3390/jintelligence7020012

Bors, D. A., Stokes, T. L., Forrin, B., & Hodder, S. L. (1999). Inspection time and intelligence: Practice, strategies, and attention. *Intelligence*, *27*, 111–129. DOI: https://doi.org/10.1016/S0160-2896 (99)00010-0

Brand, C. (1981). General intelligence and mental speed: Their relationship and development. In M. P. Friedman, J. P. Das & N. O'Connor (Eds.), *Intelligence and Learning* (pp. 589–593). Boston, MA: Springer US. DOI: https://doi.org/10.1007/978-1-4684-1083-9_56

Brand, C. R., & Deary, I. J. (1982). Intelligence and 'inspection time'. In H. J. Eysenck (Ed.), *A model for intelligence* (pp. 133–148). Berlin, Heidelberg: Springer. DOI: https://doi.org/10.1007/978-3-642-68664-1_5

Caffier, P. P., Erdmann, U., & Ullsperger, P. (2003). Experimental evaluation of eye-blink parameters as a drowsiness measure. *European Journal of Applied Physiology*, *89*, 319–325. DOI: https://doi.org/10.1007/s00421-003-0807-5

Caryl, P. G. (1994). Early event-related potentials correlate with inspection time and intelligence. *Intelligence*, *18*, 15–46. DOI: https://doi.org/10.1016/0160-2896(94)90019-1

Chaiken, S. R., & Young, R. K. (1993). Inspection time and intelligence: Attempts to eliminate the apparent movement strategy. *American Journal of Psychology*, *106*, 191–210. DOI: https://doi.org/10.2307/1423167

Chung, J. Y., Yoon, H. W., Song, M. S., & Park, H. (2006). Event related fMRI studies of voluntary and inhibited eye blinking using a time marker of EOG. *Neuroscience Letters*, *395*, 196–200. DOI: https://doi.org/10.1016/j.neulet.2005.10.094

Deary, I. J. (2000). *Looking down on human intelligence: From psychometrics to the brain* (Vol. 34). Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780198524175.003.0002

Deary, I. J. (2001). Human intelligence differences: Towards a combined experimental-differential approach. *Trends in Cognitive Sciences*, *5*, 164–170. DOI: https://doi.org/10.1016/S1364-6613(00)01623-5

Deary, I. J., Caryl, P. G., Egan, V., & Wight, D. (1989). Visual and auditory inspection time: Their inter-relationship and correlations with IQ in high ability subjects. *Personality and Individual Differences*, *10*, 525–533. DOI: https://doi.org/10.1016/0191-8869(89)90034-2

Deary, I. J., Simonotto, E., Meyer, M., Marshall, A., Marshall, I., Goddard, N., & Wardlaw, J. M. (2004). The functional anatomy of inspection time: An event-related fMRI study. *NeuroImage*, *22*, 1466–1479. DOI: https://doi.org/10.1016/j.neuroimage.2004.03.047

Deary, I. J., & Stough, C. (1996). Intelligence and inspection time: Achievements, prospects, and problems. *American Psychologist*, *51*, 599–608. DOI: https://doi.org/10.1037/0003-066X.51.6.599

De Winter, J. C. F., Gosling, S. D., & Potter, J. (2016). Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. *Psychological Methods*, *21*, 273–290. DOI: https://doi.org/10.1037/met0000079

De Winter, J. C. F., Petermeijer, S. M., Kooijman, L., & Dodou, D. (2020). *Replicating five pupillometry studies of Eckhard Hess*. Manuscript submitted for publication.

Duan, X., Dan, Z., & Shi, J. (2013). The speed of information processing of 9- to 13-year-old intellec-tually gifted children. *Psychological Reports*, *112*, 20–32. DOI: https://doi.org/10.2466/04.10.49.PR0.112.1.20-32

Egan, V. (1993). Can specific inspection time strategies be inferred from their latency? *The Irish Journal of Psychology*, *14*, 253–269. DOI: https://doi.org/10.1080/03033910.1993.10557929

Egan, V. (1994). Intelligence, inspection time and cognitive strategies. *British Journal of Psychology*, *85*, 305–315. DOI: https://doi.org/10.1111/j.2044-8295.1994.tb02526.x

Egan, V., & Deary, I. J. (1992). Are specific inspection time strategies prevented by concurrent tasks? *Intelligence*, *16*, 151–167. DOI: https://doi.org/10.1016/0160-2896(92)90002-9

Egan, V., & Deary, I. J. (1993). Does perceptual intake speed reflect intelligent use of feedback in an inspection-time task? The effect of restricted feedback. *The Journal of General Psychology*, *120*, 123–137. DOI: https://doi.org/10.1080/00221309.1993.9921188

Evans, G., & Nettelbeck, T. (1993). Inspection time: A flash mask to reduce apparent movement effects. *Personality and Individual Differences*, *15*, 91–94. DOI: https://doi.org/10.1016/0191-8869(93)90045-5

Gregory, T., Nettelbeck, T., Howard, S., & Wilson, C. (2008). Inspection time: A biomarker for cognitive decline. *Intelligence*, *36*, 664–671. DOI: https://doi.org/10.1016/j.intell.2008.03.005

Grudnik, J. L., & Kranzler, J. H. (2001). Meta-analysis of the relationship between intelligence and inspection time. *Intelligence*, *29*, 523–535. DOI: https://doi.org/10.1016/S0160-2896(01)00078-2

Hill, D., Saville, C. W. N., Kiely, S., Roberts, M. V., Boehm, S. G., Haenschel, C., & Klein, C. (2011). Early electro-cortical correlates of inspection time task performance. *Intelligence*, *39*, 370–377. DOI: https://doi.org/10.1016/j.intell.2011.06.005

Holcombe, A. O. (2009). Seeing slow and seeing fast: two limits on perception. *Trends in Cognitive Sciences*, *13*, 216–221. DOI: https://doi.org/10.1016/j.tics.2009.02.005

Hutton, U., Wilding, J., & Hudson, R. (1997). The role of attention in the relationship between inspection time and IQ in children. *Intelligence*, *24*, 445–460. DOI: https://doi.org/10.1016/S0160-2896(97)90059-3

Irwin, R. J. (1984). Inspection time and its relation to intelligence. *Intelligence*, *8*, 47–65. DOI: https://doi.org/10.1016/0160-2896(84)90006-0

Jensen, A. R. (2006). *Clocking the Mind*. Amsterdam: Elsevier.

Johns, M., Crowley, K., Chapman, R., Tucker, A., & Hocking, C. (2009). The effect of blinks and saccadic eye movements on visual reaction times. *Attention, Perception, & Psychophysics*, *71*, 783–788. DOI: https://doi.org/10.3758/APP.71.4.783

Johnson, W., & Deary, I. J. (2011). Placing inspection time, reaction time, and perceptual speed in the broader context of cognitive ability: The VPR model in the Lothian Birth Cohort 1936. *Intelligence*, *39*, 405–417. DOI: https://doi.org/10.1016/j.intell.2011.07.003

Kovacs, K., & Conway, A. R. (2016). Process overlap theory: A unified account of the general factor of intelligence. *Psychological Inquiry*, *27*, 151–177. DOI: https://doi.org/10.1080/1047840X.2016.1153946

Kranzler, J. H., & Jensen, A. R. (1989). Inspection time and intelligence: A meta-analysis. *Intelligence*, *13*, 329–347. DOI: https://doi.org/10.1016/S0160-2896(89)80006-6

Larson, G. E., & Alderton, D. L. (1990). Reaction time variability and intelligence: A "worst performance" analysis of individual differences. *Intelligence*, *14*, 309–325. DOI: https://doi.org/10.1016/0160-2896(90)90021-K

Larson, G. E., Saccuzzo, D. P., & Brown, J. (1994). Motivation: Cause or confound in information pro-cessing/intelligence correlations? *Acta Psychologica*, *85*, 25–37. DOI: https://doi.org/10.1016/0001-6918(94)90018-3

Levy, P. (1992). Inspection time and its relation to intelligence: Issues of measurement and meaning. *Personality and Individual Differences*, *13*, 987–1002. DOI: https://doi.org/10.1016/0191-8869(92)90132-9

Mackenzie, B., & Bingham, E. (1985). IQ, Inspection time, and response strategies in a university population. *Australian Journal of Psychology*, *37*, 257–268. DOI: https://doi.org/10.1080/00049538508256403

Mackenzie, B., & Cumming, S. (1986). How fragile is the relationship between inspection time and intelligence: The effects of apparent-motion cues and previous experience. *Personality and Individual Differences*, *7*, 721–729. DOI: https://doi.org/10.1016/0191-8869(86)90043-7

Miller, E. M. (1994). Intelligence and brain myelination: A hypothesis. *Personality and Individual Differences*, *17*, 803–832. DOI: https://doi.org/10.1016/0191-8869(94)90049-3

Nettelbeck, T. (2001). Correlation between inspection time and psychometric abilities. *Intelligence*, *29*, 459–474. DOI: https://doi.org/10.1016/S0160-2896(01)00072-1

Nettelbeck, T., Robson, L., Walwyn, T., Downing, A., & Jones, N. (1986). Inspection time as mental speed in mildly mentally retarded adults: Analysis of eye gaze, eye movement, and orientation. *American Journal of Mental Deficiency*, *91*, 78–91.

Nettelbeck, T., & Vita, P. (1992). Inspection time in two childhood age cohorts: A constant or a developmental function? *British Journal of Developmental Psychology*, *10*, 189–197. DOI: https://doi.org/ 10.1111/j.2044-835X.1992.tb00572.x

Nettelbeck, T., & Young, R. (1990). Inspection time and intelligence in 7-yr-old children: A follow-up. *Personality and Individual Differences*, *11*, 1283–1289. DOI: https://doi.org/10.1016/0191-8869(90)90155-K

Oberauer, K. (2019). Working memory and attention. *Journal of Cognition*, *2*, 36. DOI: https://doi.org/ 10.5334/joc.79

Posthuma, D., De Geus, E. J. C., & Boomsma, D. I. (2001). Perceptual speed and IQ are associated through common genetic factors. *Behavior Genetics*, *31*, 593–602. DOI: https://doi.org/10.1023/ A:1013349512683

Ritchie, S. J., Bates, T. C., Der, G., Starr, J. M., & Deary, I. J. (2013). Education is associated with higher later life IQ scores, but not with faster cognitive processing speed. *Psychology and Aging*, *28*, 515–521. DOI: https://doi.org/10.1037/a0030820

Salmoni, A. W., Schmidt, R. A., & Walter, C. B. (1984). Knowledge of results and motor learning: A review of critical reappraisal. *Psychological Bulletin*, *95*, 355–386. DOI: https://doi.org/10.1037//0033-2909.95.3.355

Sargezeh, B. A., Ayatollahi, A., & Daliri, M. R. (2019). Investigation of eye movement pattern parameters of individuals with different fluid intelligence. *Experimental Brain Research*, *237*, 15–28. DOI: https://doi.org/10.1007/s00221-018-5392-2

Seibel, R. (1963). Discrimination reaction time for a 1,023-alternative task. *Journal of Experimental Psychology*, *66*, 215–226. DOI: https://doi.org/10.1037/h0048914

Simpson, C. R., & Deary, I. J. (1997). Strategy use and feedback in inspection time. *Personality and Individual Differences*, *23*, 787–797. DOI: https://doi.org/10.1016/S0191-8869(97)00105-0

Spearman, C. (1923). *The nature of intelligence and the principles of cognition*. London, UK: Macmillan.

Stankov, L. (2004). Similar thoughts under different stars: Conceptions of intelligence in Australia. *International Handbook of Intelligence*, 344–363. DOI: https://doi.org/10.1017/CBO9780511616648.013

Stough, C., Bates, T. C., Mangan, G. L., & Colrain, I. (2001). Inspection time and intelligence: Further attempts to eliminate the apparent movement strategy. *Intelligence*, *29*, 219–230. DOI: https://doi.org/10.1016/S0160-2896(00)00053-2

Sullivan, E. V., Brumback, T., Tapert, S. F., Prouty, D., Fama, R., Thompson, W. K., …, & Clark, D. B. (2017). Effects of prior testing lasting a full year in NCANDA adolescents: contributions from age, sex, socioeconomic status, ethnicity, site, family history of alcohol or drug abuse, and baseline performance. *Developmental Cognitive Neuroscience*, *24*, 72–83. DOI: https://doi.org/10.1016/j.dcn.2017.01.003

Unsworth, N., Redick, T. S., Lakey, C. E., & Young, D. L. (2010). Lapses in sustained attention and their relation to executive control and fluid abilities: An individual differences investigation. *Intelligence*, *38*, 111–122. DOI: https://doi.org/10.1016/j.intell.2009.08.002

Vickers, D., Nettelbeck, T., & Willson, R. J. (1972). Perceptual indices of performance: The measurement of 'inspection time' and 'noise' in the visual system. *Perception*, *1*, 263–295. DOI: https://doi.org/10.1068/ p010263

Vickers, D., & Smith, P. L. (1986). The rationale for the inspection time index. *Personality and Individual Differences*, *7*, 609–623. DOI: https://doi.org/10.1016/0191-8869(86)90030-9

Wai, J., Lubinski, D., & Benbow, C. P. (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology*, *101*, 817–835. DOI: https://doi.org/10.1037/a0016127

White, M. (1996). Interpreting inspection time as a measure of the speed of sensory processing. *Perso nality and Individual Differences*, *20*, 351–363. DOI: https://doi.org/10.1016/0191-8869(95)00171-9

SUPPLEMENTARY MATERIALS

Table S1. First experiment: Means (*M*) and standard deviations (*SD*) of performance measures

	Response accuracy (%)		Response provided (%)		Mean response time (ms)		Median response time (ms)	
Exposure time (ms)	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
14	52.79	19.70	91.50	16.85	1305	525	1179	538
21	55.60	19.60	92.65	14.68	1267	493	1151	527
35	69.35	21.46	94.08	13.64	1189	575	1079	618
42	74.68	23.52	94.49	13.76	1133	596	1020	615
56	81.96	20.66	94.76	13.91	1083	558	974	576
83	84.80	23.89	94.29	15.62	1062	612	979	650
104	85.87	22.38	95.44	13.15	1044	591	963	621
153	86.25	22.79	95.37	12.35	1049	563	958	580

Note. The results for 'response accuracy' and 'mean response time' were based on 146 or 147 participants. The results for 'response provided' were based on 147 participants.

Table S2. Follow-up experiment: Means (*M*) and standard deviations (*SD*) of performance measures (*N* = 159)

	Response accuracy (%)		Median response time (ms)	
Exposure time (ms)	*M*	*SD*	*M*	*SD*
14	57.58	19.03	785	343
21	69.18	19.26	711	256
35	86.10	15.13	610	203
42	90.12	11.96	605	196
56	93.13	11.39	564	169
83	96.66	7.61	547	162
104	97.04	7.59	548	150
153	98.55	6.74	562	182

Note. Because of anticipatory responses, response data were unavailable for 7 of 12720 trials.

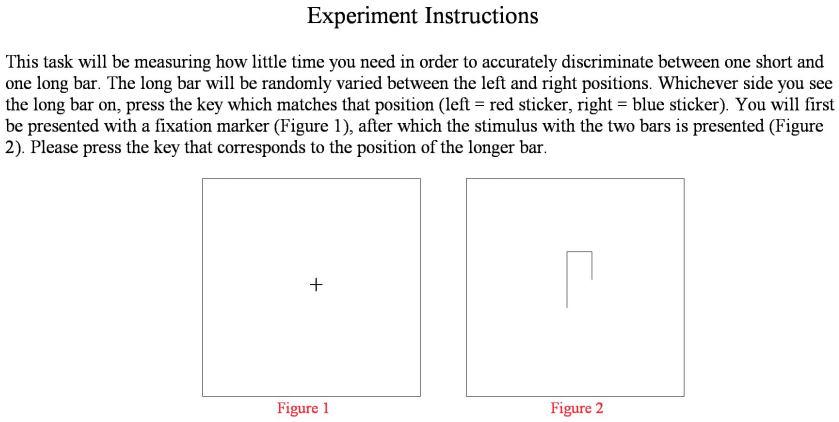


Figure S1. Instructions given prior to the experiment.

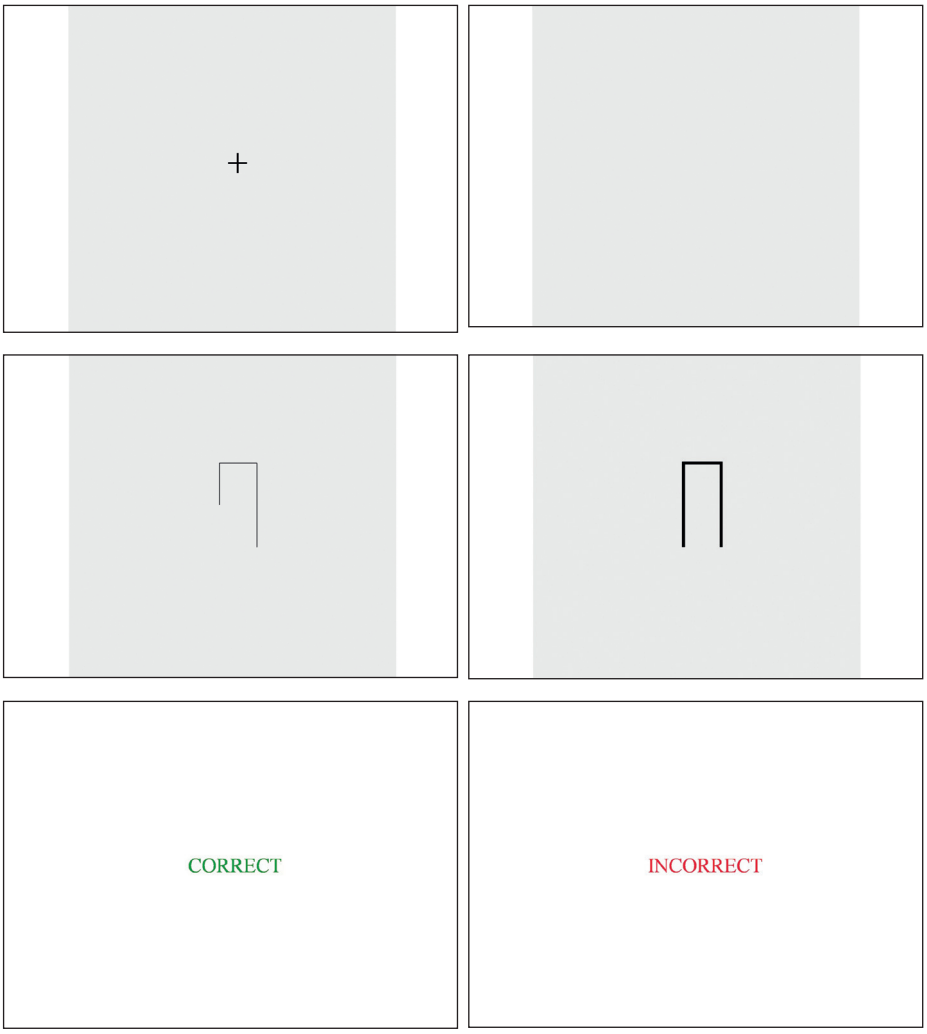


Figure S2a. First experiment: Components of an inspection time trial. Left top = Fixation marker, Right top = Blank screen, Left middle = Stimulus with one long and one short leg (i.e., the Pi-figure), Right middle = mask, Bottom left = Feedback after a correct response, Bottom right = Feedback after an incorrect response. The black figure outlines were not shown to the participants.

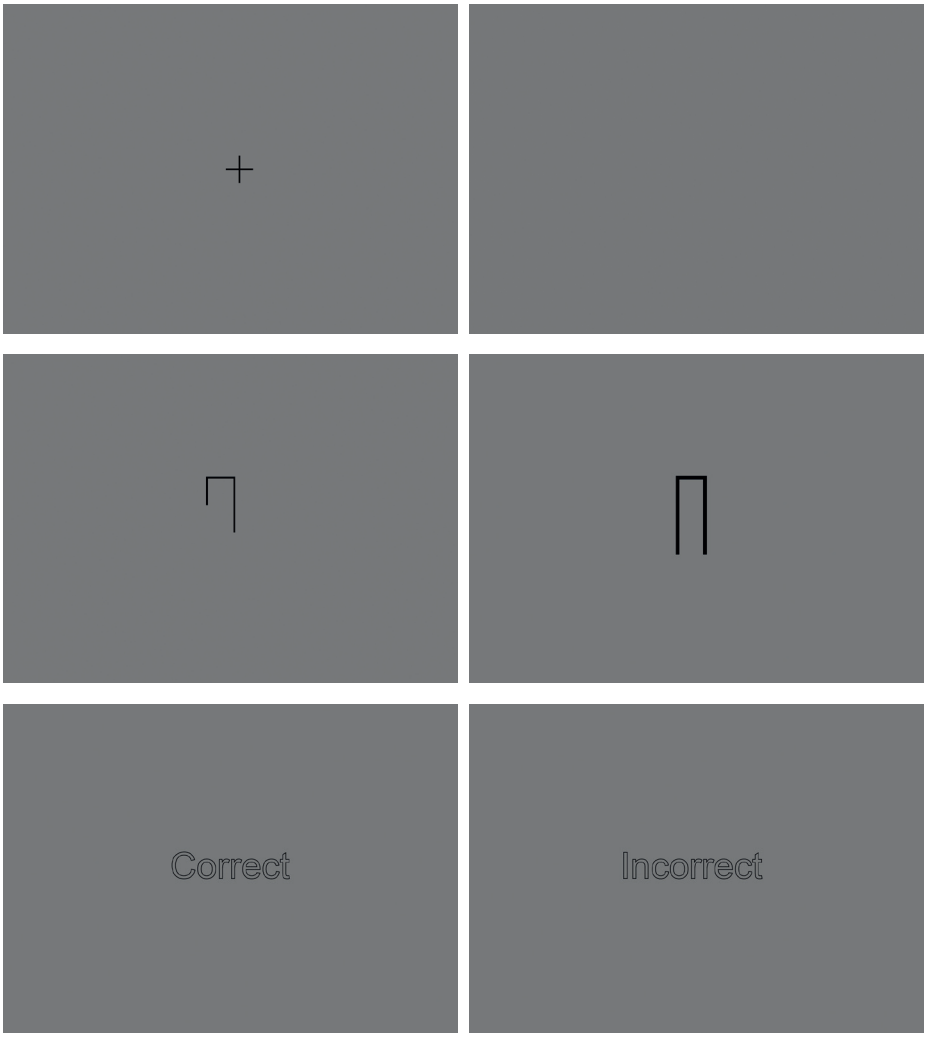


Figure S2b. Follow-up experiment: Components of an inspection time trial. Left top = Fixation marker, Right top = Blank screen, Left middle = Stimulus with one long and one short leg (i.e., the Pi-figure), Right middle = mask, Bottom left = Feedback after a correct response, Bottom right = Feedback after an incorrect response.

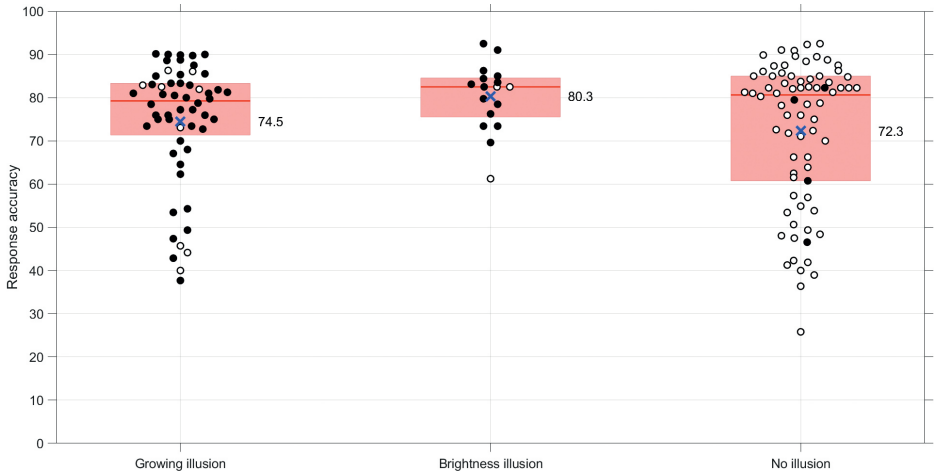


Figure S3a. First experiment: Response accuracy per participant as a function of reported illusion type. The boxplot shows the 25th percentile, median, and 75th percentile. The blue markers indicate the means. Black markers represent participants who reported 'yes' to the cue use question; white markers represent participants who reported 'no' to this question.

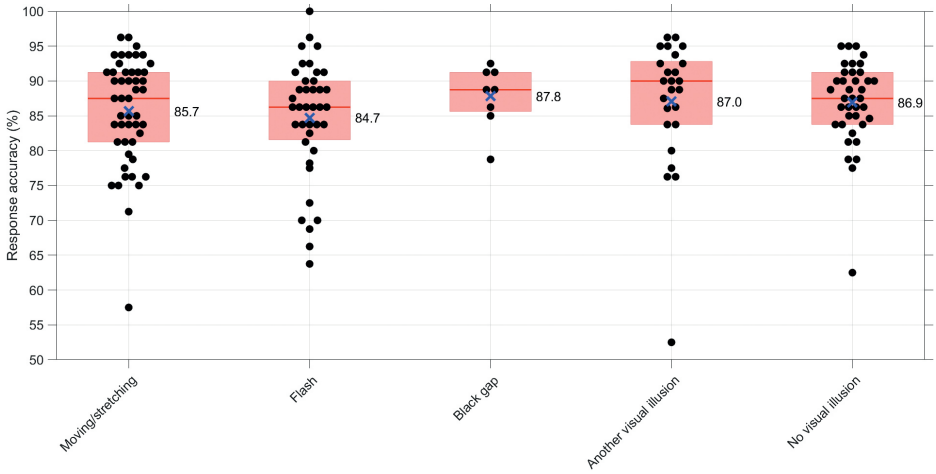


Figure S3b. Follow-up experiment: Response accuracy per participant as a function of reported illusion type. The boxplot shows the 25th percentile, median, and 75th percentile. The blue markers indicate the means.

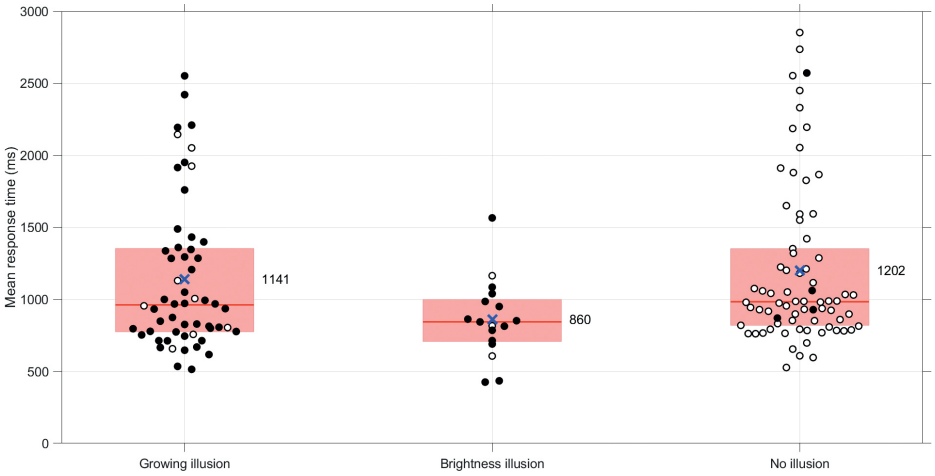


Figure S4a. First experiment: Mean response time per participant as a function of reported illusion. The boxplot shows the 25th percentile, median, and 75th percentile. The blue markers indicate the means. Black markers represent participants who reported 'yes' to the cue use question; white markers represent participants who reported 'no' to this question.

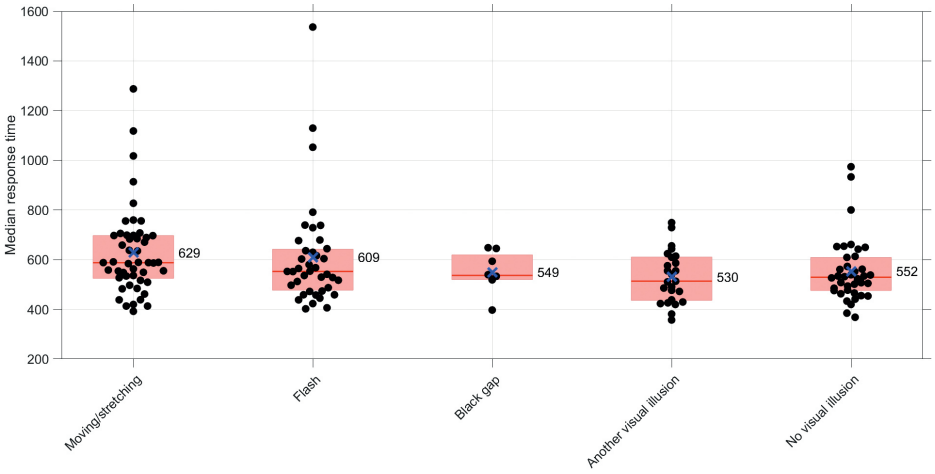


Figure S4b. Follow-up experiment: Median response time per participant as a function of reported illusion. The boxplot shows the 25th percentile, median, and 75th percentile. The blue markers indicate the means.

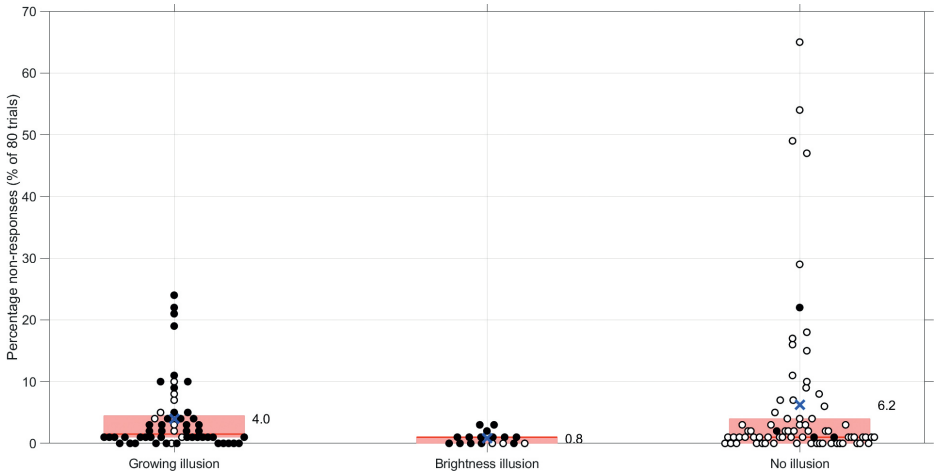


Figure S5. First experiment: Percentage of 80 trials in which the participant did not provide a response within the allocated time as a function of reported illusion type. The boxplot shows the 25th percentile, median, and 75th percentile. The blue markers indicate the mean. Black markers represent participants who reported 'yes' to the cue use question; white markers represent participants who reported 'no' to this question.

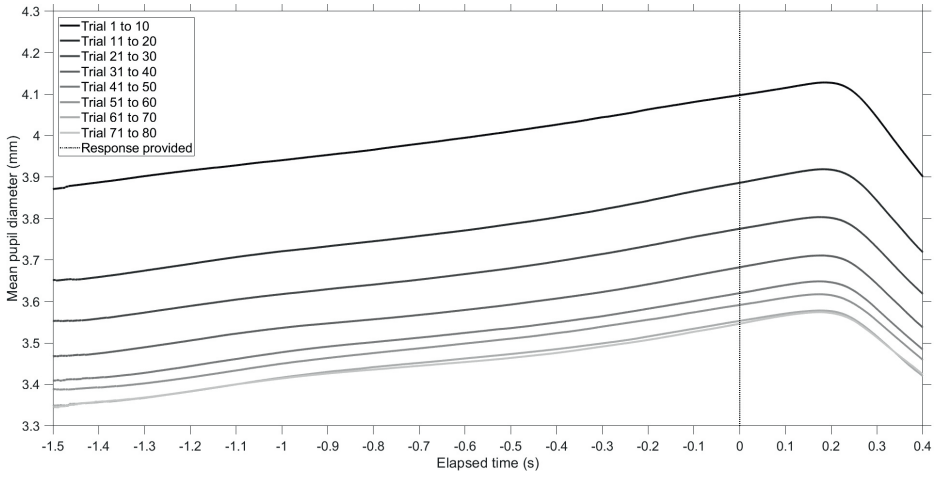


Figure S6. First experiment: Mean pupil diameter per group of 10 trials (11044 trials in total). The results are time-locked to the participants' response (occurring at t = 0 s). Participants were provided with a "CORRECT" or "INCORRECT" feedback message after responding. Pupil diameter data were linearly interpolated during blinks. Data were included up to 0.4 s after the participant provided a response.

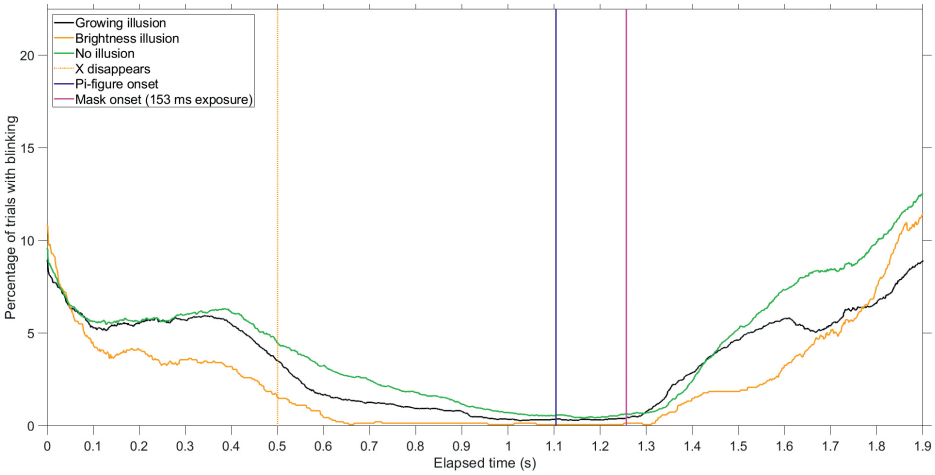


Figure S7. First experiment: Percentage of trials where the participant was blinking for the growing illusion (*n* = 4257), brightness illusion (*n* = 1345), and no illusion (*n* = 5442). Vertical lines are shown for the moment the fixation marker (X) disappeared, the moment the Pi-figure was presented, and the moment the mask was presented for the maximum exposure time of 153 ms. Data were included up to 0.4 s after the participant provided a response.

CHAPTER 10

Discussion

What drives our eyes over a visual scene, and to what extent is the distribution of visual gaze indicative of task performance? In other words: to what extent is visual attention a mediating variable between task conditions and performance? In the Introduction to this thesis, I set the goal of constructing a measure of overt visual attention as a candidate for predicting task performance. Using Just and Carpenter's (1976, 1980) 'mind-eye assumption', I assumed that the information that is present at one's position of visual gaze largely reflects the contents of one's cognitive processes. The hypothesis of this thesis is therefore that visual attention is likely to provide a good understanding of the operator's situation awareness, decision-making processes, and task performance (see also Chase & Simon, 1973). This hypothesis was evaluated in four different areas, namely: (1) an abstract monitoring task as conceived by John Senders (1964, 1983), (2) Air Traffic Control (ATC)-like tasks, (3) tasks that involve the observation of external Human Machine Interfaces (eHMIs) on automated vehicles, and (4) the allegedly simplest task in psychometrics: Inspection Time (IT). As set out in the Introduction, if applicable, Wickens's (2008) Saliency Expectance Effort Value (SEEV) model will be used as an interpretative framework for the discussion.

The results of this thesis will be summarized and discussed for each chapter individually, in light of the aforementioned aim. Finally, limitations of the present thesis will be discussed, and recommendations will be given. For each chapter, a figure highlights via dark rectangles which of the four SEEV factors (Saliency, Expectancy, Effort, or Value) are addressed in that particular chapter. Furthermore, for each chapter, its main contribution is highlighted in italics.

Chapters 2 and 3: Replicating Senders (1983)

In 1967, John Senders and colleagues published a study on visual attention in car driving, entitled: "The attentional demand of automobile driving". In this work, which incidentally won an IgNobel prize, Senders et al. (1967) proposed a mathematical model that describes visual attention distribution based on the top-down factor 'driver uncertainty'. Uncertainty pertained to the position of the own vehicle on the road, as well as the potential presence of other road users or obstacles on the road. Senders hypothesis was that human drivers sample the road in a discrete manner: they only take a sample (i.e., look at the road) when the uncertainty of the situation exceeds a certain level. Subsequent experiments confirmed this hypothesis by showing that the speed that one dares to drive depends on the number and duration of the glances the driver directs at the road. More specifically, "the less frequent the observations, or the shorter the period of observation, the slower will be the speed that the driver can maintain, and, conversely, that the greater the level at which the speed is fixed the more

often the driver must look at the road" (p. 32). Senders et al. suggested that driving performance and visual attention are intimately connected, and that glance frequency, an operationalization of visual attention, is predictive of driving performance.

Senders et al.'s (1967) applied research was preceded by a paper he wrote in 1964. Therein, he developed a more generic model of human sampling behavior, based on the Nyquist-Shannon sampling theorem (Shannon, 1948). In 1983, Senders published his thesis, in which he proposed a number of additional models that expanded on his 1964 work. These expanded models not only included top-down factors (Expectancy, per the Nyquist-Shannon theorem) but also bottom-up (Saliency) factors. Chapter 3 of my thesis provided tutorial-like in-depth explanations of Senders's (1964 and 1983) work. Senders (1983) did not report on the operator's task performance, nor on potential experimental validations of his more advanced models of visual attention distribution. Chapter 2 of the present thesis addressed these issues via a large-scale experiment in which Senders's (1983) six-dial experiment was replicated. In Chapter 3, the results were interpreted according to Wickens's (2008) SEEV model.

The main contribution of Chapter 2 was that Senders's (1964) original experimental findings regarding the relationship between signal bandwidth and glance frequency were replicated with very high accuracy (r = 0.99); that is, participants indeed sampled according to bandwidth (Expectancy). At the same time, the experiment of Chapter 2 showed that sampling behavior was not just Expectancy-driven; Effort and Saliency also proved to influence sampling distribution to a great degree. In Chapter 2, we identified three inter-related Salient features that drove participants to sample a specific dial:

1) Speed of the pointer. When the pointer moved fast, a higher proportion of participants sampled that specific dial. Speed seems, therefore, to be a Salient attractor of attention.

2) Relative pointer angle. In case the pointer came closer to a threshold, a higher percentage of participants sampled that specific dial and vice versa. The position with respect to a certain threshold (or other important parts of a scene) seems, therefore, to be a Salient attractor of attention.

3) Time-to-crossing. There was a strong relationship between the pointer's time-to-crossing and the proportion of participants who sampled that specific dial. The temporally closer the pointer was to crossing the threshold, the more participants sampled that specific dial and vice versa.

Furthermore, participants sampled in a more bandwidth-dependent manner in lower-Effort configurations, and in a less bandwidth-dependent manner in the higher-Effort configurations. Performance should, according to Chapter 3 of this thesis and Senders's own work, be a function of the participant's adherence to these 'ideal sampling' models.

The main contribution of Chapter 3 is the mathematical transparency and the illustrative graphs and figures with which Senders's (1983) 'ideal' models of visual sampling are presented and explained. The aim of Chapter 3 was to provide Human Factors researchers who are not familiar with these models and their mathematical representation with clear examples and figures as to how the models work and how they could be applied in their own work. In Chapter 2, there were indications as to the confirmation of the 'adherence' hypothesis: when the fast-moving dials were placed in the middle of the bank (i.e., a low-Effort configuration) participants performed somewhat better (i.e., detected more threshold crossings of the pointers) as compared to when the fast-moving dials were placed in the periphery (high-Effort configuration). This finding may prove useful in the design of instrument panels and cockpits.

The present evaluation of the different SEEV components as presented in Chapter 2, appears to be a novelty in the scientific literature. Christopher Wickens himself (personal communication, February 2019) indicated: "In my mind, the direct implications of Effort to SEEV have never been systematically examined in the way you did". Having replicated seminal research of Senders (1983) and interpreted the findings using the SEEV framework, the subsequent chapters studied eye movements and visual attention in different application areas (aviation, driving, psychometrics).

Chapter 4: Visual attention and Situation Awareness

The replication of Senders's (1983) six-dial experiment, as described in Chapter 2, served an additional purpose, namely to examine whether the construct of Situation Awareness (SA) can be objectively measured online (i.e., while performing a task) through eye-movements. *The main contribution of Chapter 4 is the extensive explanation regarding the construct of SA, how it is generally measured, its reported criterion validity with respect to performance, but also the construct's inherent shortcomings, and a new way of operationalizing SA through eye-movements.* The construct of SA was formalized by Mica Endsley in the 1980s (e.g., see Endsley, 1988, 1995, 2019), where she defined SA as: "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" (Endsley, 1988, p. 792). According to Endsley's (1995) model of SA, the process of gaining Situation Awareness consists of three levels. Level 1 comprises the perceptual processes, Level 2 consists of a sense-making process, and Level 3 denotes the projection of the system's future status. Furthermore, Endsley developed the Situation Awareness Global Assessment Technique (SAGAT, Endsley, 1988), an offline questionnaire that aims to measure the status of the three SA levels. Chapter 4 showed, amongst other things, that SAGAT has good criterion validity with respect to

task performance (see for an overview Endsley, 2019); correlations of up to $r = 0.65$ have been reported in the literature (Gardner, et al. 2017). However, there are also studies that show no correlation with task performance at all. Chapter 4 thereby questioned the predictive ability (and validity) of SAGAT-based SA assessments.

Moreover, SAGAT does not allow for online SA measurement, something that is imperative for operator state estimation. After all, the simulation must be stopped or frozen to administer the SAGAT. In Chapter 4, eye-movements were proposed as a real-time measure for Situation Awareness, as opposed to the discrete SAGAT method. Eye-movements (operationalized in the sampling score) were found to be highly indicative of task performance. As an addition to the replication of the six-dial experiment of Senders (1983), we administered SAGAT-like questionnaires in between the videos that depicted the dials, asking the participants to draw the last remembered position of the pointers for each of the dials. These SAGAT-like scores correlated only moderately with performance, $r = 0.20$, whereas our novel sampling score correlated highly with performance, $r = 0.78$, where the sampling score indicates the percentage of crossings for which the participants fixated on the dial within a reasonable time limit.

The results further showed that in case a participant did not sample a specific dial, the probability that the participant would press the spacebar was low (28.4%); on the other hand, if a participant sampled a dial, the probability of pressing the spacebar was high (60.8%). *In the context of the general aim of this thesis, it appears safe to say that where a person looks is more indicative of performance as compared to what a person remembers of the situation as measured by SAGAT.* Chapter 4, however, only considered foveal attention, and we could not explain why still 28.4% of the participants were able to press the spacebar correctly. This effect is possibly due to participants detecting the dials through peripheral vision, a hypothesis that needs to be tested in future research.

Chapter 5 and 6: Visual attention distribution in Air Traffic Control

Saliency	Expectancy	Effort	Value

Chapters 5 and 6 were concerned with visual attention in Air Traffic Control-like (ATC) tasks. Due to the abstract nature of the experimental stimuli and circumstances, the results of these experiments may not be directly transferable to practical recommendations for real-life ATC. Still, precisely because of this abstract nature, the reported experiments facilitate a deeper understanding of conflict-detection task performance in terms of eye-movements and the SEEV model. Apart from using the SEEV model to describe and explain performance and eye-movements, Chapter 6 allowed for evaluating the potential normative nature of the SEEV model, as was explained in the Introduction. In Chapter 5, a simple conflict-detection experiment was conducted. Participants were tasked to determine if two moving dots were about

to collide or would pass each other safely. This experiment was conceptualized in the context of conflict-detection in Air Traffic Control (cf. Loft et al., 2009; Neil & Kwantes, 2009). The main reason for the study in Chapter 5 was the lack of studies in literature that directly operationalized conflict detection performance in terms of eye-movements (cf. Hunter & Parush, 2009)

The main contribution of Chapter 5 was that eye-movements were mainly directed at the moving dots, which indicates that the majority of the eye-movements were explained by the Salient nature of the dots (most likely due to their movement, as well as the contrast with the background). Further proof of Saliency-driven sampling was provided by the fact that participants barely sampled the conflict point area of interest (AOI). For example, for the 30-degree trials only 0.2% of the gaze coordinates were directed at the conflict point AOIs, whereas 81.4% of the gaze coordinates were directed at the AOIs that surrounded the dots. *Another contribution of this study was the finding that Effort also played a role in how participants distributed their eye-movements. Dots that were farther apart invoked shorter fixations (i.e., less pursuit movement) and higher-amplitude saccades.*

In addition to eye-movement analysis, we used a novel continuous performance measurement (De Clerq, et al. 2019) in Chapter 5. Instead of asking participants to press the response key once to indicate if the dots would collide, we asked them to keep the spacebar pressed if they thought the dots would collide, and to release the spacebar if they thought otherwise. The advantage of this continuous measure is that it reflects the participant's decision-making process in time because it requires a continuous response from the participant. Participant's performance on 100-degree conflicts was lower in comparison to the other conflict angles. Especially in the case of non-conflict scenarios, it took a relatively long time (approximately 13 seconds in a 20-second trial) for the participants to realize that there was no conflict existent in the trial. The 30- and 150-degrees non-conflict scenarios showed almost identical performance; at approximately 10 seconds, most of the participants started to realize that no conflict was present. *These results point to another important contribution of this paper: the use of perceptual heuristics. The factors of the SEEV model do not fully account for conflict-detection performance; humans also make use of perceptual heuristics to perform well in a task*. *For example, it appeared plausible that for 30 degrees, participants applied the 'closer is first' heuristic (Tresilian, 1995).* After all, for two dots that travel almost parallel (and at the same speed), it is relatively easy to see if one dot trails behind the other or not. For 150-degree conflict angles, conflicts may be easy to detect, for example, by determining how far the dots are away from an imaginary straight line that connects both. However, for 100-degree conflicts, no such heuristics exist, and participants need to rely on laborious cognitive extrapolation for conflict detection, a hypothesis that appears to be supported through the displayed data in Figure 4 of Chapter 5.

In Chapter 6, a similar conflict-detection task was used. However, here we did not use moving stimuli, but employed still frames of aircraft in possible conflict. An extra between-subjects manipulation was added in the form of an augmented feedback display, called the Solution Space Diagram (SSD). *The main contribution of Chapter 6 consisted in the experimental evaluation of this SSD, mainly in terms of eye-movements and conflict detection performance.* Furthermore, instead of dots, the aircraft were represented by means of small squares (analogous to real ATC displays) and also featured a speed vector (the length corresponded to the speed of the aircraft). The participants' task was to indicate through a single spacebar press if the two displayed aircraft were in conflict, that is: whether the aircraft would collide in the near future.

Conflict detection performance was vastly different between the two groups: participants in the SSD group were approximately twice as fast in detecting conflicts, and also were more accurate. An analysis of eye-movements revealed that participants in the SSD group spent a great deal of time looking at the SSD. *These results are relatively easy explained through two cognitive mechanisms: 1) the SSD is a visually Salient feature of the scene, therefore attracting attention, which is Saliency-based sampling according to the SEEV model, and 2) the SSD provides a binary answer as to whether there will be a conflict in the future, acting on the operator's Expectancy of the situation.* In other words, the SSD attracted attention because it provided the answer to the experimental task. Also, analogous to the results in Chapter 5, Effort proved to be a mediating variable in eye-movement distribution; aircraft that were placed farther apart invoked higher saccade amplitudes.

To improve the safety of augmented feedback like the SSD, the SEEV model may be employed. In particular, the SEEV model could be used to optimize visual attention distribution and help prevent negative phenomena like cognitive tunneling. In the context of Chapter 6 for example, participants appeared to experience cognitive tunneling; that is, they spent a considerable portion of their time looking at the SSD, which came at the expense of not looking at other task-relevant features (Schmidt & Wulf, 1997). If the automation is working properly, this may not be a problem. However, in case the automation fails and the feedback is not working correctly, the operator may not be able to perform the task satisfactorily anymore. *An important finding of Chapter 6 is that the SEEV model could be used as a normative structure to guide the design of augmented feedback in such a way that potentially distracting bottom-up factors (which are Saliency and Effort) are minimized, or utilized in such a way that top-down based (Expectancy and Value) visual sampling is encouraged.* Since the SSD is a Salient feature of the visual scene and distracts visual attention at the expense of other task-relevant cues, the design could be adjusted by providing both aircraft with a visual feedback feature like the SSD. In case the automation fails, the operator is then expected to be more inclined to sample other task-relevant parts of the visual scene, rather than

focusing on the specific element that used to have the augmented feedback. To apply the SEEV model here in a quantitative way, the objective visual salience of the elements in the scene must be determined (e.g., by means of the saliency algorithms as developed by Itti & Koch, 2000). Consequently, the visual features that need (Expectancy-based) attention from the operator, could get an equal Salient appearance (e.g. 2 aircraft with the SSD), or the Saliency of the important elements could be varied, depending on the state of the system and the fixation position of the operator (e.g., gaze-contingent feedback).

Chapters 7 and 8: Visual attention distribution in AV-pedestrian interaction

In contrast to the abstract nature of the ATC tasks, Chapters 7 and 8 aimed to measure performance and visual attention in a less mundane task environment: deciding to cross or not to cross the street in front of automated vehicles (AVs). In these chapters, we researched the effect of so-called external Human Machine Interfaces (eHMIs) on participant's crossing decisions and eye-movements. In Chapter 7, a total of five eHMI positions (plus one control condition) were evaluated. A similar continuous performance measure as in Chapter 5 was used. In Chapter 8, we evaluated the effect of the eHMI's message perspective on participants crossing decisions and eye-movements. In other words, should textual eHMIs employ an allocentric (i.e., informative) message perspective, or an egocentric (i.e., instructive) message perspective?

Chapter 7 provided a good test case for the SEEV model. The eHMIs consisted of large screens that either displayed the text "Waiting" and an icon of a walking person, or "Driving" with an icon of the maximum speed. The eHMIs appeared on the roof, windscreen, grill, as a projection on the road, and above the front wheels of the car. There was also one condition without an eHMI. Participants were asked to indicate whether they felt safe to cross, by holding the spacebar pressed, and in case they did not feel safe to cross, they had to release the spacebar. Spacebar press behavior was highly contingent on which type of eHMI the car featured; generally, participants did feel safer to cross in front of a car that featured an eHMI, as compared to no eHMI at all.

The analysis of eye-movements revealed the following trends, as structured according to the SEEV model:

1) Saliency-driven sampling: the eHMIs clearly received attention as their presence was visually Salient with respect to the rest of the car. Especially during the first second of a car approaching, eye-movements were directed to the car and its eHMI.

2) Expectancy driven sampling: Eye-movements were very goal-directed; participants often appeared to focus on potential future interactions. Participants did not

necessarily sample the nearest or most Salient car but had clear expectations as to what cars would do in the near future.

3) Effort-driven sampling: eHMIs that potentially distracted the gaze from task-intrinsic cues were sampled more intermittently and created a higher gaze dispersion score. In other words, if the eHMI was placed further away from important parts of the visual scene, which was especially the case for projection-eHMI, participants started to sample laboriously between eHMI and other parts of the visual scene.

4) Value-based sampling: participants focused their attention on approaching cars as opposed to cars that were driving away, a trend that was found consistently throughout the experiment.

This finding provides to some evidence that participants Valued the cars differently within different conditions. However, this hypothesis may also be explained by the participant's Expectancy of the different AVs. One argument that would confirm Value-based sampling is that approaching cars were of objective Value to the participant's crossing decisions, and therefore shaped the participant's expectations. Cars that drove off, and therefore disappeared from the screen, were of no Value anymore to the task at hand. However, further research is needed to identify if Value is indeed a driving factor in the sampling behavior of vulnerable road users; there we could operationalize Value as an independent variable, for example in the form of a monetary reward or cost.

In Chapter 8, an experiment was conducted to investigate the effect of different eHMI message perspective on participant's response times and eye movements. In Chapter 8, we used photos of an approaching car with a textual eHMI that showed one of the following texts in a white font: WALK and DON'T WALK (egocentric message perspective), DRIVING and BRAKING (allocentric message perspective), and GO and STOP (ambiguous message perspective). This level of experimental control (photos instead of videos or interactive simulations) and performance measurement (response times instead of actual crossing behaviors) poses an advantage to more ecologically oriented studies (e.g., Cefkin et al., 2019; Kooijman et al., 2019), as it allows for evaluating the effect of nothing but the message perspective, rather than also measuring effects of so-called implicit communication (e.g., time to collision, gap size). To study the effect of mental load on participants decision-making, a memory task condition was also included.

Participants' response times and objective clarity scores were, respectively, lowest and highest for the egocentric messages. Even when the mental load was increased, egocentric messages invoked the best performance. These findings may be explained by the instructive nature of the egocentric messages: participants were literally instructed what to do. Memory load caused response times to drop, rather than the hypothesized increase but did not have a substantial effect on objective clarity scores. Furthermore,

the experiment did not reveal any significant effect of message perspective on response times per se, but rather showed that participants reacted faster to messages they experienced as clearer (e.g., GO and DRIVING, as compared to BRAKING and STOP). An eye-movement analysis revealed that longer texts instigated a higher number of saccades; a correlation of $r = 0.92$ between the subtended angle of the text and the mean number of saccades was found. In other words, longer texts need more saccades to read, a finding that is not surprising. However, text length was not indicative of task performance, even in light of higher cognitive loads.

The findings reported in Chapter 8 suggest that message perspective itself was not of influence on task performance. Rather, participants could have used heuristics-based processing in their response strategies. Especially for the memory task conditions, participants may have tried to shed tasks (i.e., responding to the eHMI stimulus and entering the remembered digits of the memory) as quickly as possible to minimize memory decay. The effect of message perspective on objective clarity scores could also be explained through the difference between the instructive and informative nature of the used messages: egocentric texts were instructive in nature, whereas allocentric messages were only informative. Consequently, the difference in objective clarity may not be attributed to the act of perspective taking per se, but due to the inherent clearness or vagueness of the message. Longer texts do not cause longer response times. Summarizing, these findings suggest that textual eHMIs are safe to be employed in real traffic, a conclusion that runs counter to expert opinion (Tabone et al., 2020).

Here, SEEV also may function as a normative structure to design safe eHMIs. For example: if one wants to attain optimal understanding and perception of the eHMI, one could equip the car with omnidirectional eHMIs (cf. Chapter 7) to capture attention (Saliency) and to facilitate sufficient understanding (Expectancy) at the same time.

Chapter 9: Visual attention distribution in Inspection Time task.

Saliency	**Expectancy**	Effort	**Value**

Finally, I examined whether visual attention can explain performance on highly standardized elementary tasks. For the last chapter in this thesis, I decided to focus on Inspection Time, purportedly the simplest psychometric tasks available.

Chapter 9 set out to identify the effect of attention on task performance in a task that is allegedly not mediated by attention-related processes at all: Inspection Time. This psychometric task was conceived in the 1970s (Vickers, Nettelbeck & Wilson, 1972) and has supposed criterion validity with respect to psychometric intelligence (Brand & Deary, 1982). Inspection time has been regarded as a direct index of mental speed, and has been popularized by intelligence researchers as being able to reflect a person's

intelligence directly. The Inspection Time task consists of a pi-like figure with two legs of unequal lengths that are connected at the top. Participants are exposed very briefly to the stimulus (typically between 2 ms and 150 ms), and have to indicate which of the legs (left of right) was longest after the stimulus has been covered by a mask. A meta-analysis on the correlation between IT and IQ reported a substantial correlation of -0.52 (Grudnik & Kranzler, 2001), apparently confirming the hypothesized predictive ability of the task. The interesting characteristic of IT tasks is that performance is supposedly not mediated by motor responses, as would be for example the case in Reaction Time (RT) tasks (Jensen, 2006), or in the words of Kranzler and Jensen (1993) "IT, the only index of mental speed that does not involve either motor (output) components or executive cognitive processes (metaprocesses), is held to tap individual differences in the 'speed of apprehension,' the quickness of the brain to react to external stimuli prior to any conscious thought." (pp. 329–330). In summary, IT performance should only be affected by perceptual processes and not be contaminated by motor-activity.

In light of the aim of this thesis (i.e., to research whether attention is a mediating variable between task conditions and performance), it is interesting to further investigate task performance in the Inspection Time task and operationalize the effect of focused attention. Attention was operationalized as 'not blinking', differing from the more direct operationalizations (i.e., eye-movements) as used in the other Chapters of this thesis. The main reason for this being the short time-frames in which the stimulus appeared, thus, providing only a very limited time-frame in which to acquire eye-tracking data. Also, as IT should be free of any motor effects, the suppression of blinks (which is tied to motor activity) should not be of influence on task performance.

In Chapter 9, we identified the following main effects that indicate that the IT task is not as simple as has been previously believed:

1) The temporal placement of blinks correlated with task performance. Participants who blinked after stimulus presentation performed better at the task. This effect could be seen as a manifestation of Expectancy-based sampling. In other words: participants had learned when to blink based on their expectations as to when the stimulus would appear. These findings were supported by the observation of clear learning curves.

2) Understanding the task is a likely predictor of how well people performed the task. In experiment 1 of Chapter 9, many participants indicated that they did not understand the task, In Study 1, the response accuracy therefore topped at around 75%, whereas in Study 2, in which new participants were provided with more extensive task instructions, the response accuracy went up to approximately 85%. In terms of the SEEV model, we could interpret the correct understanding of the task as being Value-driven. Participants who did not understand the task, were not

attributing enough Value to the pi-figure of the task (which is necessary for correct performance).

3) Perceptual heuristics play a large role in how participants performed the task. In Study 2, we asked participants to reflect on their usage of higher-level cognitive strategies. The results of these reflections revealed that participants employed the perception of different visual illusions, a phenomenon we believe to be an epiphenomenon of good task performance.

In Chapter 9, the role of Expectancy is striking: participants knew, or had learned, when to blink. However, as in Chapter 5, many participants used perceptual or cognitive heuristics to perform the task; in case of the IT task, these heuristics consisted of the utilization of visual illusions in the context of higher-order cognitive strategies. This finding supports our observation from Chapter 5: the SEEV model may have to be expanded with perceptual heuristics in order to able to predict task performance accurately.

Conclusion

In the Introduction to the thesis, I set out to construct a measure of visual attention that would be predictive of task performance. Did I succeed in reaching this goal?

1) **Yes, the task conditions, as qualitatively classified using SEEV, influence gaze behavior**. **The chapters in this thesis have shown that each of the variables (Expectancy, Value, Saliency, Effort) influence the probability that an operator will glance towards an object in the visual scene.** For example, Saliency proved to be a strong predictor of visual attention; Salient features of the visual scene often attracted much more attention as compared to lesser Salient features. In case of Chapter 2, faster moving dials (which are more Salient to the eye) attracted more attention as compared to the slower moving dials. Also, Expectancy seemed to be an excellent predictor of where visual attention would reside. In Chapter 2, higher bandwidth dials often attracted more attention, simply because participants expected that more events would be taking place there. Participants directed their attention to specifically those cars which they expected to be of relevance to the task at hand.

2) **Yes, gaze behavior is predictive of task performance**. In simple tasks, such as monitoring dials, strong predictive validity could be obtained (r = 0.8), which appears to be stronger than the predictive validity of a freeze-probe method. However, it must be noted that looking at something does not necessarily mean that one also sees it. Just and Carpenter's (1980) 'immediacy hypothesis' turns out to be quite a strong assumption; the fact that gaze behavior has no higher criterion validity (r = 0.8 as compared with r = 1) as reported is partly due to this 'look-but-

not-see' phenomenon (see the discussion in Chapter 4) but may also attributed to peripheral sampling. Future research that is being carried out by the author is aimed at identifying the exact contribution of peripheral vision on visual scanning behavior and consequently task performance.

Furthermore, eye-movements revealed to which extent operators used visual feedback (SSD, eHMI), which in turn improved task performance. However, in other tasks, eye-movements did not predict task performance very accurate; for instance, for the tasks that involved predicting future conflicts (i.e. SA levels 2 & 3) and for tasks that were of very short-lasting nature (rapid decision making for eHMIs). Even for the shortest-nature task (IT), some predictive validity of eye-movement analysis was obtained. Here, the participant's attention allocation was found to predict performance on one of the simplest/fastest tasks that can be imagined.

Eye-movements in itself are not entirely predictive performance, but their measurement facilitates an objective insight in the constituents of task performance. This cautionary notion on the predictive ability of visual attention is confirmed by the results as described, for example, in Chapter 5. Here, eye-movements were not predictive of performance per se, but rather provided cues as to how people performed on the task on a more conceptual level, for example by suggesting the participants' use of perceptual heuristics. The same observation applies to the results of Chapter 9, which showed that task heuristics are predictive of task performance.

3) **Considering that the task conditions influence eye-movements to a very strong extent, and considering that eye-movements predict task performance to some extent, eye-movements appear to be a mediating variable between the task and performance.** This was the central thesis of Chapter 3. I conclude that any future real-time SA and task-performance predictor should connect task conditions and eye-movements. In particular, a computational model shall define where an operator should look and compare this to the operator's actual looking behavior. This thesis may have laid the first step towards real-time assessment (e.g., see Chapters 3 and 4 in particular).

4) **Apart from real-time performance prediction, this thesis also showed that eye-movements allow for normative assessments about task- and human-machine interface design**. For example, based on eye-movement analyses, it was concluded that eHMIs should not be projected on the road, that augmented feedback should not be an attention-grabber that distracts from relevant elements of the scene, and that text-based eHMIs are not as damaging as experts may currently believe (Tabone et al., 2020).

5) Measuring eye-movements on itself will not necessarily teach us anything about task performance. By using the SEEV model as a framework to interpret the eye-movement data, we were able to discern the different bottom-up and top-down factors at play in the different application areas. SEEV does, of course, not account for all factors that determine visual attention distribution. For example, the SEEV model does not yet account for the use of perceptual heuristics, which proved to be important for explaining sampling behavior and task performance in some of the studies in this thesis. To conclude: **SEEV model is indeed an excellent tool for interpreting eye-movements, and it provides a semantic structure for explaining visual sampling behavior in context of a task. However, the use of heuristics should be accounted for in the SEEV model, in order to increase its predictive validity even more.**

REFERENCES

Brand, C. R., & Deary, I. J. (1982). Intelligence and 'Inspection Time'. *A Model for Intelligence,* 133-148. doi:10.1007/978-3-642-68664-1_5

Cefkin, M., Zhang, J., Stayton, E., & Vinkhuyzen, E. (2019). Multi-methods research to examine external HMI for highly automated vehicles. *HCI in Mobility, Transport, and Automotive Systems Lecture Notes in Computer Science,* 46-64. doi:10.1007/978-3-030-22666-4_4

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4*(1), 55-81. doi:10.1016/0010-0285(73)90004-2

Clercq, K. D., Dietrich, A., Velasco, J. P., Winter, J. C. F., & Happee, R. (2019). External human-machine interfaces on automated vehicles: effects on pedestrian crossing decisions. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 61*(8), 1353-1370. doi:10.1177/0018720819836343

Endsley, M. (1988). Situation awareness global assessment technique (SAGAT). *Proceedings of the IEEE 1988 National Aerospace and Electronics Conference*. doi:10.1109/naecon.1988.195097

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 37*(1), 32-64. doi:10.1518/001872095779049543

Endsley, M. R. (2019). A systematic review and meta-analysis of direct objective measures of situation awareness: a comparison of SAGAT and SPAM. *Human Factors: The Journal of the Human Factors and Ergonomics Society,* 001872081987537. doi:10.1177/0018720819875376

Gardner, A. K., Kosemund, M., & Martinez, J. (2017). Examining the feasibility and predictive validity of the SAGAT Tool to assess situation awareness among medical trainees. *Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare,* 1. doi:10.1097/sih.0000000000000181

Gibson, J. J., & Gibson, E. J. (1955). Perceptual learning: differentiation or enrichment? *Psychological Review, 62*(1), 32-41. doi:10.1037/h0048826

Grudnik, J. L., & Kranzler, J. H. (2001). Meta-analysis of the relationship between intelligence and inspection time. *Intelligence, 29*(6), 523-535. doi:10.1016/s0160-2896(01)00078-2

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40*(10-12), 1489-1506. doi:10.1016/s0042-6989(99)00163-7

Jensen, A. R. (2006). *Clocking the mind: Mental chronometry and individual differences*. Amsterdam: Elsevier.

Just, M. A., & Carpenter, P. A. (1976). The role of eye-fixation research in cognitive psychology. *Behavior Research Methods & Instrumentation, 8*(2), 139-143. doi:10.3758/bf03201761

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87*(4), 329-354. doi:10.1037/0033-295x.87.4.329

Kooijman, L., Happee, R., & Winter, J. D. (2019). How do eHMIs affect pedestrians' crossing behavior? a study using a head-mounted display combined with a motion suit. *Information, 10*(12), 386. doi:10.3390/info10120386

Kranzler, J. H., & Jensen, A. R. (1993). Psychometric g is still not unitary after eliminating supposed "impurities": Further comment on Carroll. *Intelligence, 17*(1), 11-14. doi:10.1016/0160-2896(93)90033-2

Loft, S., Bolland, S., Humphreys, M. S., & Neal, A. (2009). A theory and model of conflict detection in air traffic control: Incorporating environmental constraints. *Journal of Experimental Psychology: Applied, 15*(2), 106-124. doi:10.1037/a0016118

Neal, A., & Kwantes, P. J. (2009). An evidence accumulation model for conflict detection performance in a simulated air traffic control task. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 51*(2), 164-180. doi:10.1177/0018720809335071

Schmidt, R. A., & Wulf, G. (1997). Continuous concurrent feedback degrades skill learning: implications for training and simulation. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 39*(4), 509-525. doi:10.1518/001872097778667979

Senders, J. (1964). The human operator as a monitor and controller of multidegree of freedom systems. *IEEE Transactions on Human Factors in Electronics, HFE-5*(1), 2-5. doi:10.1109/thfe.1964.231647

Senders, J. W. (1983). *Visual sampling processes* (Unpublished master's thesis). Tilburg.

Senders, J., Kristofferson, A., Levison, W., Dietrich, C., & Ward, J. (1967). The attentional demand of automobile driving. *Highway Research Record, 195*, 15-33.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*(4), 623-656. doi:10.1002/j.1538-7305.1948.tb00917.x

Thomas, L., & Wickens, C. (2006). Display dimensionality, conflict geometry, and time pressure effects on conflict detection and resolution performance using cockpit displays of traffic information. *The International Journal of Aviation Psychology, 16*(3), 321-342. doi:10.1207/s15327108ijap1603_5

Todd, A. R., Cameron, C. D., & Simpson, A. J. (2017). Dissociating processes underlying level-1 visual perspective taking in adults. *Cognition, 159*, 97-101. doi:10.1016/j.cognition.2016.11.010

Tresilian, J. R. (1995). Perceptual and cognitive processes in time-to-contact estimation: Analysis of prediction-motion and relative judgment tasks. *Perception & Psychophysics, 57*(2), 231-245. doi:10.3758/bf03206510

Vickers, D., Nettelbeck, T., & Willson, R. J. (1972). Perceptual indices of performance: the measurement of 'Inspection Time' and 'Noise' in the visual system. *Perception, 1*(3), 263-295. doi:10.1068/p010263

Vlakveld, W. (2011). *Hazard anticipation of young novice drivers: Assessing and enhancing the capabilities of young novice drivers to anticipate latent hazards in road and traffic situations*. Leidschendam: Stichting Wetenschappelijk Onderzoek Verkeersveiligheid.

Wickens, C., & McCarley, J. (2008). *Applied Attention Theory*. Boco-Ratan, FL: CRC Press, Taylor & Francis.

Winter, J. C. (2014). Controversy in human factors constructs and the explosive use of the NASA-TLX: A measurement perspective. *Cognition, Technology & Work, 16*(3), 289-297. doi:10.1007/s10111-014-0275-1

Winter, J. C. F., Eisma, Y. B., Cabrall, C. D., Hancock, P. A., & Stanton, N. A. (2018). Situation awareness based on eye movements in relation to the task environment. *Cognition, Technology & Work, 21*(1), 99-111. doi:10.1007/s10111-018-0527-6

APPENDICES

Nawoord

Dankwoord

List of Publications

Curriculum vitae

NAWOORD

"Wege, nicht Werke" (Martin Heidegger, 1889 – 1976)

Lectori salutem!

Dit nawoord richt zich tot de enkeling die ik, met Kierkegaard, mijn lezer noem. Vermoedelijk leest u dit nawoord als eerste, en verwordt het daarmee op een eigenaardige manier tot een voorwoord. Vermoedelijk leest u alleen dit nawoord, en verwordt het daarmee op een eigenaardige manier tot de kern van dit geschrift. Wat u ook doet, laat ik in elk geval diegenen die dit nawoord lezen, met Kierkegaard een woord ter waarschuwing toeschikken: "vergeet […] toch niet dat het een dubbelzinnige kunst is te kunnen spreken, en zelfs dat het een heel twijfelachtige volmaaktheid is het ware te kunnen zeggen".

Het werk wat hier achter mij ligt heeft primair de intentie om de in de experimenten verkregen waarnemingen uit te drukken in een technische taal, te vervatten in modellen en vergelijkingen, en dat te presenteren in het strakke stramien van het wetenschappelijk artikel. Dit proefschrift – een gestructureerd exposé van experimentele resultaten, grafieken en vergelijkingen – heeft als doel om de waarheid aan het licht te brengen. Ik bedoel daarmee te zeggen: te zien of mijn verwachtingen van de wereld, hypothesen genaamd, een kern van waarheid bevatten. Dat betekent: of datgene wat uitgesproken wordt in die hypothesen overeenkomt met de waarnemingen zoals gedaan in het experiment. De waarheid die dit proefschrift dientengevolge aan het licht brengt, fungeert dus louter in de sfeer van de 'adequatio', als overeenstemming van een hypothese (of zo u wilt: construct) met de waargenomen zaak. Behandel dit proefschrift dan ook als zodanig, als iets wat een beperkte zeggingskracht in het pragmatische heeft, maar verder moet zwijgen.

Kierkegaard schrijft ergens in een voorwoord: "dit boek wil slechts zijn wat het is, iets overbodigs". Lees daarom dit proefschrift op die wijze zoals Wittgenstein zijn Tractatus Logicus Philosophicus eindigt:

"6.54 Mijn zinnen verhelderen daardoor, dat hij, die mij begrijpt, ze uiteindelijk als onzinnig herkent, wanneer hij door hen – op hen – over hen naar boven geklommen is. (Hij moet zogezegd de ladder wegwerpen, nadat hij erop naar boven geklommen is.) Hij moet deze zinnen overwinnen, dan ziet hij de wereld juist.

7. Waarvan men niet spreken kan, daarover moet men zwijgen."

In dat bewustzijn gaat dit boekje de wereld in.

DANKWOORD

Allereerst wil ik Joost de Winter, mijn dagelijks begeleider en mentor tijdens het PhD-traject, bedanken voor alle gedane moeite en zorgen. Zonder zijn uitmuntende ondersteuning en begeleiding was er vermoedelijk weinig van mijn academische pretenties en wensen terecht gekomen. Ik bedank ook Clark Borst, Rene van Paassen en Max Mulder voor de begeleiding en support in het eerste gedeelte van mijn PhD-traject. Ook mijn collega's van de afdeling Cognitive Robotics met wie ik verschillende kamers gedeeld hen: hartelijk dank voor alle collegialiteit en samenwerking.

Aan mijn vrouw Gea en mijn beide jongens Bauke en Sytze, denk ik met liefde. Met hen begin ik elke dag vol hoop en liefde. Zij leren mij elke dag de les van de lelie en de vogel: zuivere dankbaarheid en ontvankelijkheid. Al begint de dag grijs, de zon gloort altijd aan de horizon als in een eeuwige dageraad.

Mijn familie dank ik ook voor alle ondersteuning, op mentaal en materieel vlak. Mijn ouders, Bauke en Baukje Eisma, voor hun mentale ondersteuning als het ging over de weg naar de academie. Jerke, voor zijn voorbeeldfunctie; een excellent wetenschapper en een bijzondere broeder. Jacob, Sijke, Alie, Jacoba en Eelco, dank voor jullie hartelijke verbondenheid.

Mijn vrienden dank ik voor de sociale ondersteuning en voor de gezelligheid rondom het door ons menigmaal geheven glas. In het bijzonder Aart. We hebben samen boeken gelezen, filosofische discussies gevoerd en artikelen geschreven en daar zijn we hopelijk nog lang niet mee klaar. Ivo, bedankt voor alle gesprekken en vriendschap de afgelopen jaren. Gerralt, dank voor de vele keren dat we elkaar van hart tot hart spraken.

Rest mij om iedereen te bedanken die ik hier niet genoemd heb, of vergeten ben te noemen. SDG.

Yke Bauke Eisma

Sleeuwijk, Maart 2021.

LIST OF PUBLICATIONS

The author of this thesis independently supervised the MSc and BSc students involved in the separate chapters (Chapters 5, 7, 8). Two chapters involve a joint first authorship (Chapters 2 and 3).

Chapter 2. Eisma, Y. B., Cabrall, C. D. D., & De Winter, J. C. F. (2018). Visual sampling processes revisited: Replicating and extending Senders (1983) using modern eye-tracking equipment. *IEEE Transactions on Human Machine Systems, 48,* 526–540.

Chapter 3. De Winter, J. C. F., Eisma, Y. B., Cabrall, C. D. D., Hancock, P. A., & Stanton, N. A. (2019). Situation awareness based on eye movements in relation to the task environment. *Cognition, Technology and Work, 21,* 99–111.

Chapters 2 and 3 are an extension of the author's MSc thesis entitled: Eisma. Y. B. (2017). *Visual sampling and situation awareness.* Delft University of Technology, the Netherlands.

Chapter 4. Eisma, Y. B., Hancock, P. A., & De Winter, J. C. F. (in press). On Senders's models of visual sampling behavior. *Human Factors*.

Chapter 5. Eisma, Y. B., Looijestijn, A. E., & De Winter, J. C. F. (2020). Attention distribution while detecting conflicts between converging objects: An eye-tracking study. *Vision, 4,* 34.

Chapter 6. Eisma, Y. B., Borst, C., Van Paassen, M. M., & De Winter, J. C. F. (in press). Augmented visual feedback: Cure or distraction? *Human Factors.*

Chapter 7. Eisma, Y. B., Van Bergen, S., Ter Brake, S. M., Hensen, M. T. T., Tempelaar, W. J., & De Winter, J. C. F. (2020). External human-machine interfaces: The effect of display location on crossing intentions and eye movements. *Information, 11*, 13.

Chapter 8. Eisma, Y. B., Reiff, A. Kooijman, L., Dodou, D., & De Winter, J. C. F. (in press). External human-machine interfaces: Effects of message perspective. *Transportation Research Part F*, 78, 30–41.

Chapter 9. Eisma, Y. B., & De Winter, J. C. F. (2020). How do people perform an Inspection Time task? An examination of illusions, learning, and blinking. *Journal of Cognition, 3,* 34.

Papers not included in this dissertation

Bazilinskyy, P., & Dodou, D., Eisma, Y. B., Vlakveld, W. V. , & De Winter, J. C. F. (2020). Blinded windows and empty driver seats: The effects of automated vehicle characteristics on cyclist decision-making. Manuscript submitted for publication.

Bazilinskyy, P., Eisma. Y. B., Dodou, D., & De Winter, J. C. F. (2020). Risk perception: A study using dashcam videos and participants from different world regions. *Traffic Injury Prevention, 21,* 347–353.

De Winter, J. C. F., Dodou, D., Happee, R. & Eisma, Y. B. (2019). Will vehicle data be shared to address the how, where, and who of traffic accidents? *European Journal of Futures Research, 7,* 2.

Kaleefathullah, A. A., Merat, N., Lee, Y. M., Eisma, Y. B., Madigan, R., Garcia, J., & De Winter, J. C. F. (in press). External Human-Machine Interfaces can be misleading: An examination of trust development and misuse in a CAVE-based pedestrian simulation environment. *Human Factors.*

List of open datasets

Eisma, Y. B., Reiff, A. Kooijman, L., Dodou, D., & De Winter, J. C. F. (2021). Supplementary materials for the article: External human-machine interfaces: Effects of message perspective. 4TU.Centre for Research Data [Dataset]. https://doi.org/10.4121/14068553.v1

Bazilinskyy, P., Eisma, Y. B., Dodou, D., & De Winter, J. C. F. (2020). Supplementary materials for the article: Risk perception - A study using dashcam videos and participants from different world regions. *4TU.Centre for Research Data* [Dataset]. https://data.4tu.nl/repository/uuid:cd649413-c707-4469-8c47-2e20a0ee1f87

Eisma, Y. B., & De Winter, J. C. F. (2020). Supplementary materials for the article: Augmented visual feedback: Cure or distraction? *4TU.Centre for Research Data* [Dataset]. https://data.4tu.nl/repository/uuid:f689c7d5-c1f4-44e3-9897-581da590ff90

Eisma, Y. B., Van Bergen, S., Ter Brake, S. M., Hensen, M. T. T., Tempelaar, W. J., De Winter, J. C. F. (2019). Supplementary materials for the article: 'External Human–Machine Interfaces: The Effect of Display Location on Crossing Intentions and Eye Movements'. *4TU.Centre for Research Data* [Dataset]. https://doi.org/10.4121/uuid:22ade94a-77af-451f-afd4-88b4ed2a36b3

De Winter, J. C. F., Eisma, Y. B., & Cabrall, C. D. D. (2018). Supplementary data for the article: Situation awareness based on eye movements in relation to the task

environment. *4TU.Centre for Research Data* [Dataset]. https://doi.org/10.4121/uuid:15d436f4-dbb2-468f-b66b-6d134ea2821c

Eisma, Y. B., Cabrall, C. D. D., & De Winter, J. C. F. (2018). Supplementary data for the article: Visual sampling processes revisited: Replicating and extending Senders (1983) using modern eye-tracking equipment. *4TU.Centre for Research Data* [Dataset]. https://doi.org/10.4121/uuid:63affb79-d408-4f5b-9b79-8238dd42fa76

CURRICULUM VITAE

Education

2017 to 2021	Ph.D. in Mechanical Engineering *Delft University of Technology* Thesis title: Visual Attention in Human-Machine Interaction Promotors: Dr. ir. J.C.F. de Winter, Dr. ir. M.M. van Paassen
2014 to 2017	M.Sc. in Mechanical Engineering *Delft University of Technology* Thesis title: Visual Sampling and Situation Awareness
2010 to 2014	B.Sc. in Mechanical Engineering *Delft University of Technology* Thesis title: The skill of bicycle riding
2007 to 2010	Voorbereidend Wetenschappelijk Onderwijs (VWO) *Dockinga College, Dokkum*
2004 to 2007	Middelbaar Algemeen Voortgezet Onderwijs (MAVO) *Mavo de Saad, Damwoude*