

**Toward early markers for  
Autism Spectrum Disorder  
using eye tracking**

Roy S. Hessels

ISBN 978-90-827020-0-2

Cover layout: Jelle Mooi & MJ Zeegers

Photography by: Ivar Pel

Printed by: Ridderprint — [www.ridderprint.nl](http://www.ridderprint.nl)

© 2017 Roy S. Hessels

**Toward early markers for Autism Spectrum Disorder using eye tracking**

**Naar vroege voorspellers van Autisme Spectrum Stoornis met oogbewegingsmetingen**

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. G.J. van der Zwaan, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op vrijdag 7 juli 2017 des middags te 4.15 uur

door

**Roy Sebastiaan Hessels**

geboren op 12 april 1990 te Voorburg

**Promotor:** Prof.dr. C. Kemner

**Copromotor:** Dr. I.T.C. Hooge

“tomorrow we will run faster, stretch out our  
arms farther.... And then one fine morning—  
So we beat on, boats against the current, borne  
back ceaselessly into the past.”

---

*F. Scott Fitzgerald*



# Contents

<b>1. Introduction</b>	<b>3</b>
<b>I. Advances in eye tracking methodology</b>	<b>23</b>
2. Qualitative tests of remote eye-tracker recovery and performance during head rotation	25
3. Consequences of eye color, positioning, and head movement for eye-tracking data quality in infant research	55
4. Noise-robust fixation detection in eye-movement data – Identification by two-means clustering (I2MC)	101
5. The area-of-interest problem in eye-tracking research: a noise-robust solution for face and sparse stimuli	157
<b>II. Visual search in ASD</b>	<b>205</b>
6. Is there a limit to the superiority of individuals with ASD in visual search?	207
7. An in-depth look at saccadic search in infancy	233
<b>III. Gaze behavior to faces in interaction</b>	<b>265</b>
8. Gaze behavior to faces during dyadic interaction	267

*Contents*

<b>9. Eye contact takes two – capturing gaze behavior of subclinical autism and social anxiety in dyadic interaction</b>	<b>313</b>
<b>10. Discussion</b>	<b>347</b>
<b>11. Samenvatting in het Nederlands</b>	<b>367</b>
<b>12. Acknowledgements</b>	<b>387</b>
<b>13. List of publications</b>	<b>395</b>
<b>14. Curriculum vitae</b>	<b>399</b>

# 1. Introduction

## 1. Introduction

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterized by “*persistent deficits in social communication and social interaction*”, and “*restricted, repetitive patterns of behavior, interests, or activities*” (Diagnostic and Statistical Manual of Mental Disorders 5 (DSM-5); American Psychiatric Association, 2013)<sup>1</sup>. Its estimated prevalence is around 1% in U.S. and non-U.S. countries (American Psychiatric Association, 2013), and this is expected to be similar in the Netherlands (Gezondheidsraad, 2009). Typical examples of the deficits in social communication and interaction are failure to initiate or respond to bids for social interaction, failure of normal turn-taking in conversation, abnormal non-verbal communication such as in making eye contact, and problems with developing and maintaining relationships. Typical examples of restricted, repetitive behavioral patterns, interests or activities are stereotyped motor movements, echolalia, inflexibility in deviating from routines and hypo- or hyperreactivity to sensory input (American Psychiatric Association, 2013). Sensory processing is suggested to be more generally affected in ASD, as individuals with ASD typically excel on certain visuospatial tasks as compared to typically-developing individuals (for reviews, see Dakin & Frith, 2005; Simmons et al., 2009).

According to Dutch clinical estimates, the average age at which ASD is diagnosed internationally is around 6 to 7 years (van Berckelaer-Onnes et al., 2015). However, several studies have reported that diagnosis of ASD may occur consistently earlier than 6 to 7 years with the advance of new screening tests (Dietz et al., 2006; Swinkels et al., 2006; Oosterling et al., 2010). Early diagnosis is more generally possible when ASD symptoms are severe (Wiggins, Baio, & Rice, 2006). Some studies report that diagnosis of ASD is reliable even under the age of 3 years (Moore & Goodson, 2003;

---

<sup>1</sup>In the previous classification (DSM-IV), ASD was part of a larger branch of pervasive developmental disorders, and separate diagnostic criteria existed for Autistic Disorder, Asperger’s Disorder, and Pervasive Developmental Disorder Not Otherwise Specified (PDD-NOS; American Psychiatric Association 2000). Although parts of this dissertation still contain diagnoses according to the DSM-IV criteria, the key diagnostic criteria remain highly similar in the DSM-5 (American Psychiatric Association, 2013), and the latter will therefore be used as guideline.

Stone et al., 1999). More recent work suggests, however, that diagnosis under 30 months of age may be much less reliable than previously estimated (Turner & Stone, 2007). A main benefit of early detection of ASD may be early intervention. Although the evidence is still scarce, parent-directed (Oono, Honey, & McConachie, 2013) and behavioral interventions (Reichow et al., 2012) for children with ASD younger than 7 years have appeared to be effective in alleviating ASD severity. A recent intervention in toddlers with ASD has shown potential in improving cognitive and adaptive functioning (Dawson et al., 2010), and the question beckons whether early interventions may generally be more effective than later interventions, and whether interventions before toddlerhood are possible. As such, large-scale projects (e.g. the British Autism Study of Infant Sibling; BASIS, and European Autism Interventions; EU-AIMS) are now being undertaken in order to establish early signs of ASD by following infants who already have an older sibling diagnosed with ASD. As ASD has a large hereditary component, there is a higher occurrence of ASD in this group. Such infants are referred to as infants at-risk for ASD.

Previous research revealed that children who are later diagnosed with ASD can retrospectively be distinguished from typically-developing (TD) children on several behavioral signs. Children with ASD looked less in the direction of faces of others and demonstrated less directed vocalizations compared to TD children at 12 months of age. At 18 months of age, children with ASD additionally showed less social smiles, and engaged less socially compared to TD children. Interestingly, at 6 months of age no differences between children with ASD and TD children were observed (Ozonoff et al., 2010). Whereas assessments based on direct examiner ratings or video coding of behavior reveal no differences between children with ASD and TD children, recently, efforts have been directed at investigating more fine-grained behavior by studying infant gaze behavior using objective techniques (e.g. Jones & Klin, 2013). The present dissertation is concerned with the investigation of two possible early markers of ASD: 1) visual search superiority and 2) gaze behavior during face perception. Visual search superiority and gaze behavior during face perception are explored as differences

## 1. Introduction

in these domains have already been observed between individuals with ASD and TD individuals at an older age. Moreover, face perception is one aspect of the social deficits in ASD, whereas visual search superiority is one aspect of the visual peculiarities in ASD, thereby investigating aspects of both diagnostic criteria for ASD.

The following sections deal first with visual search in ASD and subsequently with gaze behavior in face perception in ASD. Hereafter, the problems of eye tracking in infant research are discussed. Finally, an outline of this dissertation is presented.

### 1.1. Visual search in ASD

When one searches for an object of interest in a visual scene that cannot be located at first glance, one must consecutively inspect different areas of the visual scene by making saccades – fast, ballistic eye movements bringing another part of the visual scene into central vision. At central vision, the retina is most sensitive to small details and color, which are more difficult to resolve with increasing eccentricity. This behavior of looking at different areas of the visual scene when looking for an object of interest is referred to as visual search. Typically, a visual search task is used in which a participant is asked to determine if and where a target – defined by a feature or a combination of features – is present among a number of non-targets. When the target is defined by one feature, for example a tilted line among vertical line non-targets, it is often referred to as a *featural* search task. When the target is defined by a combination of features, for example an upright letter *T* among tilted *T* and upright *L* non-targets, it is often referred to as a *conjunction* search task (Wolfe, 1998b). Conjunction targets generally take longer to locate compared to featural targets. This is, however, not a strict rule and exceptions may occur where conjunction targets are easier to locate than featural targets. Figure 1.1 depicts typical examples of featural and conjunction search tasks.

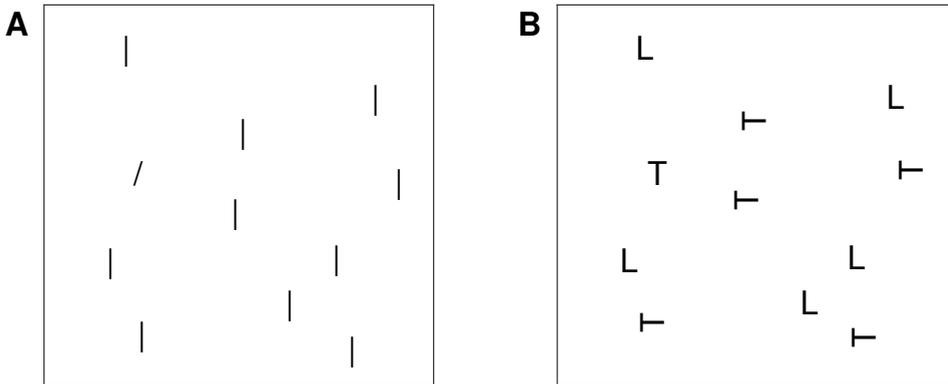


Figure 1.1.: Typical examples of visual search displays. (A) A featural search display where the target is a tilted line among vertical non-targets. (B) A conjunction search display where the upright letter T is the target, among tilted T and upright L non-targets.

In 1998, Plaisted, O’Riordan, & Baron-Cohen reported that children with ASD aged 7 to 10 years were faster at detecting a conjunctive, but not a featural target compared to TD children, without making more mistakes (i.e. there was no speed-accuracy trade-off). This finding was replicated a number of years later by O’Riordan et al. (2001), who reported that children with ASD were superior (faster without making more mistakes) at detecting a conjunctive, but not a featural target compared to TD children. Moreover, when the featural task was made more difficult by increasing the similarity between the target and the non-targets, children with ASD were also superior compared to TD children. As work on visual search in adults suggests that the dichotomy of featural versus conjunction search is not one of search mechanism, but merely of stimulus characteristics (Duncan & Humphreys, 1989; Wolfe, 1998a), O’Riordan et al. (2001) hypothesize that children with ASD outperform TD children on difficult visual search tasks (which can be either conjunction or feature search tasks). Notably, children with ASD and TD children were matched on non-verbal reasoning ability, which precluded general ability as an explanatory factor for the difference in search performance between the two groups.

## 1. Introduction

Over the years, visual search superiority in ASD has been well replicated (O’Riordan, 2004; Jarrold et al., 2005; Kemner et al., 2008; Joseph et al., 2009; Kaldy et al., 2011; Collignon et al., 2013; Keehn & Joseph, 2016), although there are also some contradicting findings (Keehn et al., 2008; Constable et al., 2010; Van Eylen et al., 2015). An overview of the current literature on visual search in ASD is depicted in Figure 1.2. Three major points can be distilled from this overview. First, visual search superiority in ASD has been observed across a wide age range. Second, visual search superiority in ASD is generally more apparent for conjunction or difficult featural targets compared to featural targets (i.e. without a specific indication as to its difficulty). Third, and most important, visual search superiority in ASD has been observed in children as early as 2.5 years. Given that visual search superiority has been reported for children with ASD in toddlerhood, the question beckons whether this is also apparent prior to the time at which an ASD diagnosis is given, perhaps even as early as infancy.

### 1.1.1. Visual search superiority as an early marker for ASD

When visual search tasks are administered to older children and adults, participants can be instructed about what the target looks like and asked to press a button when the target is found. It is subsequently easy to determine the time to and accuracy of target localization. As young children (toddlers and infants) do not generally comprehend or follow instructions, this procedure does not work for young children. How then to investigate visual search behavior in such young children? One solution is to adopt eye-tracking technology and measure where young children look on a visual search display. Nonetheless, using eye-tracking technology to study visual search in young children poses several theoretical and practical problems. One practical problem is that while much is known about eye-tracking methods and achieving high data quality with adult participants (Holmqvist et al., 2011), much less is known about eye-tracking methods and data quality in infancy. One theoretical problem is how to conclude that infants search based on their eye movements. These issues are ad-

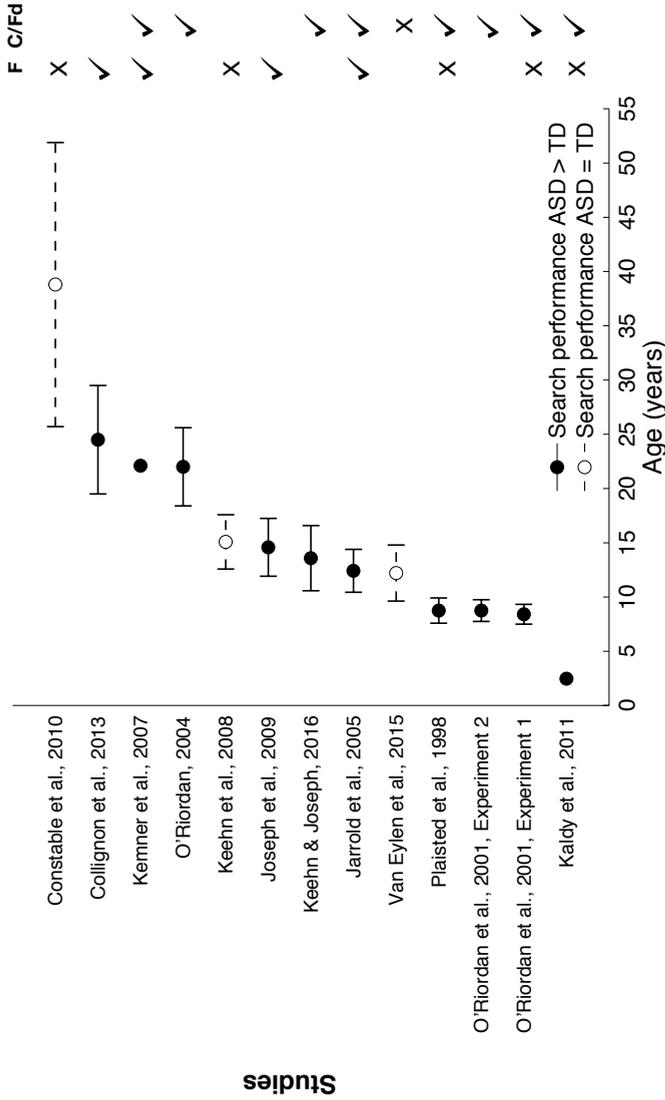


Figure 1.2.: Studies investigating visual search in ASD as a function of age of the ASD group. Solid lines indicate studies that reported visual search superiority for the ASD group in at least one of the conditions, whereas dashed lines indicate studies that did not find a difference in visual search performance in any condition between the ASD and TD group. Checkmarks on the right side denote visual search superiority for feature (F) or conjunction/difficult feature (C/Fd) targets, and crosses denote the absence of superiority for that target type.

## 1. Introduction

dressed in this dissertation.

### 1.2. Gaze behavior during face perception in ASD

One of the exemplar deficits in social communication and interaction characteristic for ASD is abnormal eye contact (American Psychiatric Association, 2013). In 2002, Pelphrey et al. were the first to use eye-tracking methodology to investigate how exactly individuals with ASD scan faces. They reported that adults with ASD looked less at the core features (eyes, nose and mouth) in photographs of faces compared to TD adults, and more at other areas of the screen. This was for the most part due to the fact that adults with ASD spent less time looking at the eyes compared to TD adults. This finding of reduced time spent looking at eye region as a model for abnormality in making eye contact seems promising. Particularly if the amount of time spent looking at the eyes may serve as a measurable index of difficulty in social interaction. However, subsequent research has proved inconsistent in reporting reduced looking time to the eyes in ASD. Several studies reported that adults (Sterling et al., 2008; Hernandez et al., 2009), adolescents (Klin et al., 2002; Dalton et al., 2005), school-aged children (Rice et al., 2012) and toddlers (Jones et al., 2008) with ASD spend less time looking at the eyes compared to TD individuals. Moreover, 6-month-old infants at risk for ASD looked less at the eyes of their mother when that mother maintained a still face halfway through a period of normal interaction compared to infants who were not at risk for ASD (Merin et al., 2007). Jones & Klin (2013) further reported that infants at-risk for ASD who later on received a diagnosis of ASD did not look less at the eyes compared to TD infants in the first 6 months after birth, but showed diminished looking at the eyes hereafter. There are, however, also reports that adults (Rutherford & Towns, 2008) and children (Van der Geest et al., 2002; Dapretto et al., 2006; McPartland et al., 2011) with ASD did not look less at the eyes compared to TD individuals. Chawarska & Shic (2009) furthermore reported that 4-year-olds but not 2-year olds with ASD looked less at the core features of a face, but no differences on time spent looking at the eyes

## 1.2. Gaze behavior during face perception in ASD

were reported. Concluding, the evidence for reduced looking time to the eyes in ASD is inconsistent.

Inconsistencies are not only reported for the time individuals with ASD spend looking at the eye region. For example, toddlers (Jones et al., 2008) and children (Klin et al., 2002) with ASD were reported to look longer at the mouth region compared to TD individuals. In the latter study, the amount of time spent looking at the mouth was furthermore positively correlated with social adaptation. Other studies however reported that individuals with ASD spent less time looking at the mouth than TD individuals (Rice et al., 2012; Chawarska & Shic, 2009). Rice et al. (2012) furthermore reported that the time spent looking at the mouth region was differentially related to ASD severity (as measured using the Autism Diagnostic Observation Schedule; Lord et al. 1989) depending on the IQ profile of the individuals with ASD.

Several explanations have been posited for the inconsistencies in gaze behavior to faces in ASD. Bird et al. (2011) posited that the reduced looking at the eyes in ASD may be explained by the presence of alexithymia in some individuals with ASD – a phenomenon characterized by difficulties in identifying and describing emotions in the self, which generally affects 50% of the individuals with ASD. Another explanation posited by Speer et al. (2007) is that the social context of the stimuli used may be very important. In their study photographs and videos of either one or more persons were employed. Individuals with ASD differed only from TD individuals in the social dynamic (videos of more people) condition, in which reduced looking time to the eyes was observed. While the social context may appear to reconcile differences in looking time to the eyes across studies, this does not reconcile differences in increased looking time to the mouth (Klin et al., 2002; Rice et al., 2012). A recent review on this topic concludes that the assumption that *“individuals with ASD exhibit excess mouth and diminished eye gaze compared to TD individuals”* (p. 286) receives little support (Guillon et al., 2014). Senju & Johnson (2009) conclude that *“the available evidence at present suggests that the reduced fixation on the eyes*

## 1. Introduction

*in ASD is most prominent under conditions of high cognitive demand* (p. 1211).

Recently, a new approach to studying behavior in social settings has been proposed, termed “Cognitive ethology” (Smilek et al., 2006; Kingstone et al., 2008; Kingstone, 2009). This approach advocates studying behavior in natural settings before moving to the laboratory instead of vice versa. Given that the social context (Speer et al., 2007; Guillon et al., 2014), or cognitive demand (Senju & Johnson, 2009), appears to be a factor in whether gaze behavior to faces in ASD is affected, new methods for investigating gaze behavior to faces in social interaction might prove promising. Moreover, given that ASD is marked by problems in social communication and interaction, it makes sense to investigate ASD symptomatology in actual interaction. In this dissertation, the Cognitive ethology approach is applied to the investigation of gaze behavior to faces in ASD. Specifically, gaze to faces during interaction is investigated as a possible avenue towards a marker for ASD.

### 1.3. Eye tracking in infant research

As has briefly been alluded to in Section 1.1, using eye-tracking technology with infant participants to study gaze behavior may not be as straightforward as with adult participants. Although technological advances in eye tracking have made it possible to study infant eye movements without placing equipment on the infant, a number of key problems remain (Aslin & McMurray, 2004). First of all, adult participants are generally cooperative and easy to instruct. This means that they can be instructed to sit directly in front of the eye tracker and remain so. Moreover, they can be asked to perform certain actions to calibrate the eye tracker (i.e. map certain recorded features of the participants’ eye to known locations on a screen) and check the accuracy of the recording. Infants, however, cannot be instructed to do so and often tend to move when movement is possible. This means that positioning an infant prior to the start of the

experiment is difficult, as well as maintaining the same positioning throughout the experiment. Moreover, calibration of an eye tracker cannot be done by asking the infant to look at a given point on screen, as the infant will not follow the instruction. Although there is some discussion on the effect of these problems on eye-tracking data quality (e.g. Wass et al., 2013), little is written about possible solutions for such problems. This holds for whether an eye-tracker can cope with certain positioning and head and/or body movement, whether and how data quality is affected by movement, and which analysis tools are suited for infant eye-tracking data. As the aim of the current dissertation is to investigate early markers of ASD using eye-tracking technology, a critical analysis of eye-tracking methodology in infant research is necessary. Such an analysis is made in the first part of the dissertation.

## **1.4. Outline of this dissertation**

The dissertation is divided into three parts. Part 1 is concerned with the advances necessary to investigate the possible early markers of ASD using eye-tracking methodology. In this part, the following questions are addressed: How does one choose an eye-tracker for infant research? How does one attain high quality eye-tracking data with infants? How can infant eye-tracking data be analysed such that data of varying quality is automatically processed and mapped to a visual stimulus? These questions are critical if one is to draw conclusions on (a)typical development based on eye-tracking data. Part 2 is concerned with visual search in ASD and the possibility of using it as an early marker. In this part, the following questions are posed: Is there a limit to the superiority of visual search in ASD, and what may be the scope? Do infants show spontaneous visual search behavior, and how can this be characterized? Knowledge on the limit and scope of visual search superiority, and of typical visual search behavior in infancy are vital if atypical visual search behavior in infancy is to be pinpointed. Part 3 is concerned with gaze behavior during face perception in ASD, and more specifically with gaze behavior during dyadic interaction (interaction

## 1. Introduction

involving two people). The questions that are addressed in this part are: Do the findings on biases in gaze to faces generalize from research using static stimuli to dyadic interaction? Is gaze to faces in dyadic interaction related to (sub)clinical ASD? Answering these questions will prove to be a first step towards investigating (a)typical gaze in dyadic interaction in infancy. Figure 1.3 depicts a schematic overview of the dissertation and its chapters.

Investigating possible early markers of ASD using eye-tracking methodology requires that an eye tracker is opted for. The choice for which eye tracker to use places restrictions on the subsequent research carried out. However, little is known about which eye trackers are suited for infant research, and which selection criteria are relevant for infant eye-tracking research. In **Chapter 2**, we outline which criteria one may use to choose an eye tracker and investigate the performance of a range of eye trackers in a simulation of behavior typically encountered in eye-tracking research with infants.

After an eye tracker is opted for, the next step is to ensure that the quality of eye-movement data obtained with infant participants is sufficient for drawing conclusions on gaze behavior. In **Chapter 3**, we investigate which factors contribute to data quality in infant eye-tracking research, and discuss means of improving data quality.

As infant eye-tracking data are generally of lower quality than adult eye-tracking data, analysis tools must be chosen such that they can deal with low-quality eye-tracking data. However, the few tools that have been built for this purpose either exclude a large proportion of eye-tracking data or require manual coding. In **Chapter 4**, we introduce an algorithm that automatically processes eye-tracking data, and is able to cope with eye-tracking data of varying noise and data loss levels.

In most eye-tracking research where the intention is to map gaze of an observer to parts of the visual stimulus, an Area-of-Interest (AOI) analysis

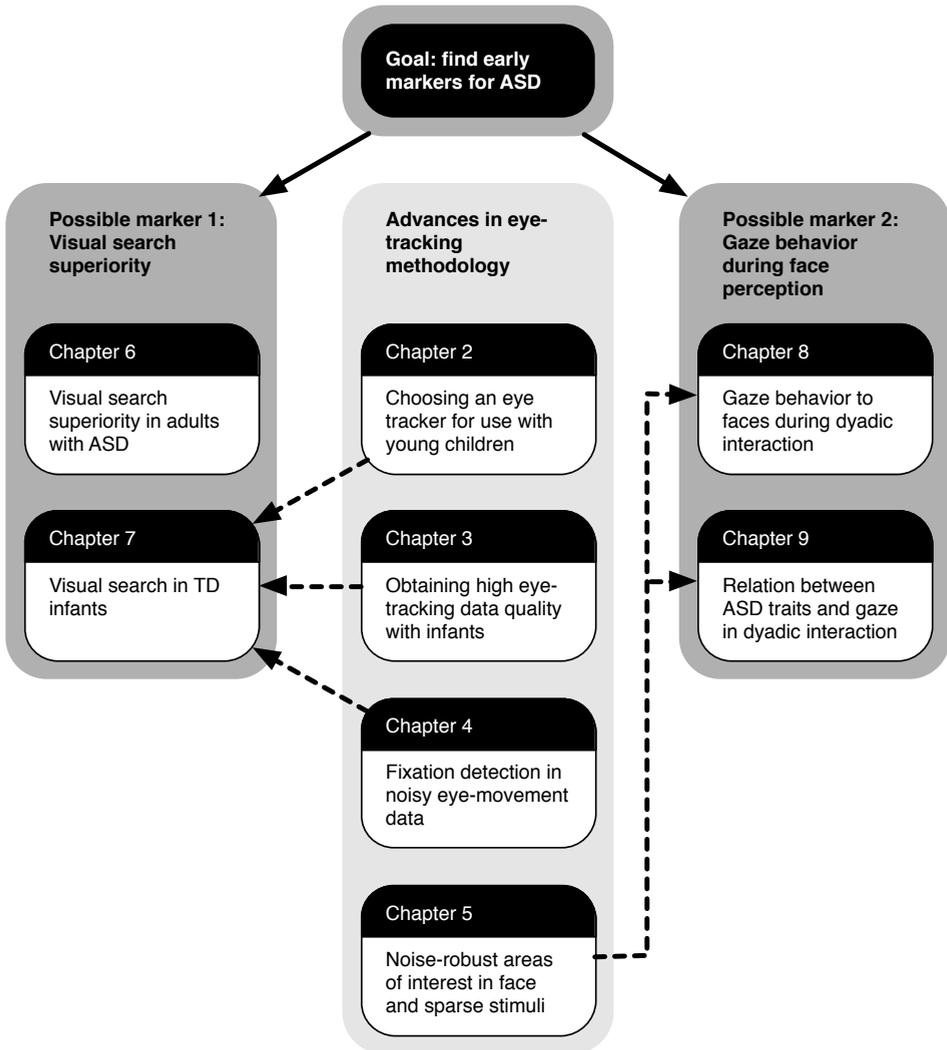


Figure 1.3.: Schematic overview of the dissertation, and its three parts. The necessity of the advances in eye-tracking methodology for the other chapters are depicted by dashed arrows.

## 1. Introduction

is used. Using this analysis, AOIs are defined for parts of the scene, and gaze is subsequently assigned to the AOIs by a computer. In **Chapter 5**, we outline the methods of AOI construction in face stimuli that are at a researcher's disposal, discuss differences between these methods in whether they can be implemented by a machine, and investigate their ability to cope with eye-tracking data of low quality.

In **Chapter 6**, we investigate visual search in adults with and without ASD, and discuss the visual search superiority that has often been reported for individuals with ASD. Moreover, we discuss whether there is a limit to the superiority of individuals with ASD in visual search, and its possible explanations.

Before visual search behavior of infants at-risk for ASD is to be investigated, knowledge is required on what constitutes typical visual search behavior across development. In **Chapter 7**, we investigate visual search in 10-month-old infants. Specific attention is paid to how one can conclude that infants search based on their eye movements. Moreover, differences and similarities are discussed between adult and infant visual search behavior.

As previous research on gaze behavior during face perception in ASD has revealed numerous inconsistencies, a new approach to gaze behavior in ASD is investigated. In **Chapter 8**, we discuss the importance of social presence and social context for the study of gaze behavior to faces, and introduce a novel dual eye-tracking social interaction setup for the investigation of gaze behavior to faces in dyadic interaction. Using this setup, we investigate whether the preference for looking at the eyes generalizes from research using static pictures to social interaction. Moreover, we investigate the relation between the gaze behavior of two observers when they have the possibility to interact.

In **Chapter 9**, the setup introduced in the previous chapter is employed to investigate the relation between ASD traits and Social Anxiety Disorder

(SAD) traits and gaze behavior to faces during dyadic interaction. Specifically, we investigate whether the previously reported (albeit inconsistent) reduced looking time at the eyes for individuals with ASD occurs for individuals scoring high on ASD traits in dyadic interaction as well. Moreover, we investigate the relation between ASD and SAD traits of both observers in a pair and their paired gaze behavior.

Finally, in **Chapter 10**, the advances in eye-tracking methodology introduced in chapters 2 to 5, the studies on visual search in ASD in chapters 6 and 7, and the studies on gaze behavior to faces in dyadic interaction and its relation to ASD in chapters 8 and 9 are discussed. Conclusions will be drawn with regard to the state of the art of eye-tracking methodology in developmental psychology, and directions for the investigation of visual search and gaze behavior during face perception as early markers for ASD will be outlined.

## 1. Introduction

## References

- American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders*. American Psychiatric Association, Arlington, VA, 4th edition.
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders*. American Psychiatric Association, Washington, DC, 5th edition.
- Aslin, R. N. & McMurray, B. (2004). Automated corneal-reflection eye tracking in infancy: Methodological developments and applications to cognition. *Infancy*, 6(2):155–163.
- Bird, G., Press, C., & Richardson, D. C. (2011). The role of alexithymia in reduced eye-fixation in autism spectrum conditions. *Journal of Autism and Developmental Disorders*, 41(11):1556–1564.
- Chawarska, K. & Shic, F. (2009). Looking but not seeing: Atypical visual scanning and recognition of faces in 2 and 4-year-old children with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, 39(12):1663–1672.
- Collignon, O., Charbonneau, G., Peters, F., Nassim, M., Lassonde, M., Lepore, F., Mottron, L., & Bertone, A. (2013). Reduced multisensory facilitation in persons with autism. *Cortex*, 49(6):1704–1710.
- Constable, P., Solomon, J. A., Gaigg, S. B., & Bowler, D. M. (2010). Crowding and visual search in high functioning adults with autism spectrum disorder. *Clinical Optometry*, 2:93–103.
- Dakin, S. & Frith, U. (2005). Vagaries of visual perception in autism. *Neuron*, 48(3):497–507.
- Dalton, K. M., Nacewicz, B. M., Johnstone, T., Schaefer, H. S., Gernsbacher, M. A., Goldsmith, H. H., Alexander, A. L., & Davidson, R. J. (2005). Gaze fixation and the neural circuitry of face processing in autism. *Nature neuroscience*, 8(4):519–526.
- Dapretto, M., Davies, M. S., Pfeifer, J. H., Scott, A. A., Sigman, M., Bookheimer, S. Y., & Iacoboni, M. (2006). Understanding emotions in others: Mirror neuron dysfunction in children with autism spectrum disorders. *Nature neuroscience*, 9(1):28–30.
- Dawson, G., Rogers, S., Munson, J., Smith, M., Winter, J., Greenson, J., Donaldson, A., & Varley, J. (2010). Randomized, controlled trial of an intervention for toddlers with autism: The early start Denver model. *Pediatrics*, 125(1): e17–e23.
- Dietz, C., Swinkels, S., van Daalen, E., van Engeland, H., & Buitelaar, J. K.

## 1.4. Outline of this dissertation

- (2006). Screening for autistic spectrum disorder in children aged 14–15 months. II: Population screening with the early screening of autistic traits questionnaire (ESAT). Design and general findings. *Journal of Autism and Developmental Disorders*, 36(6):713–722.
- Duncan, J. & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review*, 96(3):433–458.
- Gezondheidsraad. (2009). *Autismespectrumstoornissen: een leven lang anders*. Gezondheidsraad, Den Haag.
- Guillon, Q., Hadjikhani, N., Baduel, S., & Rogé, B. (2014). Visual social attention in autism spectrum disorder: Insights from eye tracking studies. *Neuroscience & Biobehavioral Reviews*, 42:279–297.
- Hernandez, N., Metzger, A., Magné, R., Bonnet-Brilhault, F., Roux, S., Barthelemy, C., & Martineau, J. (2009). Exploration of core features of a human face by healthy and autistic adults analyzed by visual scanning. *Neuropsychologia*, 47:1004–1012.
- Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press.
- Jarrold, C., Gilchrist, I. D., & Bender, A. (2005). Embedded figures detection in autism and typical development: Preliminary evidence of a double dissociation in relationships with visual search. *Developmental Science*, 8(4): 344–351.
- Jones, W. & Klin, A. (2013). Attention to eyes is present but in decline in 2–6-month-old infants later diagnosed with autism. *Nature*, 504:427–431.
- Jones, W., Carr, K., & Klin, A. (2008). Absence of preferential looking to the eyes of approaching adults predicts level of social disability in 2-year-old toddlers with autism spectrum disorder. *Archives of General Psychiatry*, 65(8):946–954.
- Joseph, R. M., Keehn, B., Connolly, C., Wolfe, J. M., & Horowitz, T. S. (2009). Why is visual search superior in autism spectrum disorder? *Developmental Science*, 12(6):1083–1096.
- Kaldy, Z., Kraper, C., Carter, A. S., & Blaser, E. (2011). Toddlers with autism spectrum disorder are more successful at visual search than typically developing toddlers. *Developmental Science*, 14(5):980–988.
- Keehn, B. & Joseph, R. M. (2016). Exploring what’s missing: What do target absent trials reveal about autism search superiority? *Journal of Autism and Developmental Disorders*.

## 1. Introduction

- Keehn, B., Brenner, L., Palmer, E., Lincoln, A. J., & Müller, R.-A. (2008). Functional brain organization for visual search in ASD. *Journal of the International Neuropsychological Society*, 14:990–1003.
- Kemner, C., Ewijk, L., Engeland, H., & Hooge, I. (2008). Brief report: Eye movements during visual search tasks indicate enhanced stimulus discriminability in subjects with PDD. *Journal of Autism and Developmental Disorders*, 38(3):553–557.
- Kingstone, A. (2009). Taking a real look at social attention. *Current Opinion in Neurobiology*, 19:52–56.
- Kingstone, A., Smilek, D., & Eastwood, J. D. (2008). Cognitive ethology: A new approach for studying human cognition. *British Journal of Psychology*, 99(3):317–340.
- Klin, A., Jones, W., Schultz, R., Volkmar, F., & Cohen, D. (2002). Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Archives of General Psychiatry*, 59:809–816.
- Lord, C., Rutter, M., Goode, S., Heemsbergen, J., Jordan, H., Mawhood, L., & Schopler, E. (1989). Autism diagnostic observation schedule: A standardized observation of communicative and social behavior. *Journal of Autism and Developmental Disorders*, 19(2):185–212.
- McPartland, J. C., Webb, S. J., Keehn, B., & Dawson, G. (2011). Patterns of visual attention to faces and objects in autism spectrum disorder. *Journal of Autism and Developmental Disorders*, 41(2):148–157.
- Merin, N., Young, G. S., Ozonoff, S., & Rogers, S. J. (2007). Visual fixation patterns during reciprocal social interaction distinguish a subgroup of 6-month-old infants at-risk for autism from comparison infants. *Journal of Autism and Developmental Disorders*, 37:108–121.
- Moore, V. & Goodson, S. (2003). How well does early diagnosis of autism stand the test of time? *Autism*, 7(1):47–63.
- Oono, I. P., Honey, E. J., & McConachie, H. (2013). Parent-mediated early intervention for young children with autism spectrum disorders (ASD). *Evidence-Based Child Health: A Cochrane Review Journal*, 8(6):2380–2479.
- Oosterling, I. J., Wensing, M., Swinkels, S. H., van der Gaag, R. J., Visser, J. C., Woudenberg, T., Minderaa, R., Steenhuis, M.-P., & Buitelaar, J. K. (2010). Advancing early detection of autism spectrum disorder by applying an integrated two-stage screening approach. *Journal of Child Psychology and Psychiatry*, 51(3):250–258.

## 1.4. Outline of this dissertation

- O’Riordan, M. A. (2004). Superior visual search in adults with autism. *Autism*, 8(3):229–248.
- O’Riordan, M. A., Plaisted, K. C., Driver, J., & Baron-Cohen, S. (2001). Superior visual search in autism. *Journal of Experimental Psychology: Human Perception and Performance*, 27(3):719–730.
- Ozonoff, S., Iosif, A.-M., Baguio, F., Cook, I. C., Hill, M. M., Hutman, T., Rogers, S. J., Rozga, A., Sangha, S., Sigman, M., Steinfeld, M. B., & Young, G. S. (2010). A prospective study of the emergence of early behavioral signs of autism. *Journal of the American Academy of Child & Adolescent Psychiatry*, 49(3):256–266.e2.
- Pelphrey, K. A., Sasson, N. J., Reznick, J. S., Paul, G., Goldman, B. D., & Piven, J. (2002). Visual scanning of faces in autism. *Journal of Autism and Developmental Disorders*, 32(4):249–261.
- Plaisted, K., O’Riordan, M., & Baron-Cohen, S. (1998). Enhanced visual search for a conjunctive target in autism: A research note. *Journal of Child Psychology and Psychiatry*, 39(5):777–783.
- Reichow, B., Barton, E. E., Boyd, B. A., & Hume, K. (2012). Early intensive behavioral intervention (EIBI) for young children with autism spectrum disorders (ASD). *Cochrane Database of Systematic Reviews*, 10.
- Rice, K., Moriuchi, J. M., Jones, W., & Klin, A. (2012). Parsing heterogeneity in autism spectrum disorders: Visual scanning of dynamic social scenes in school-aged children. *Journal of the American Academy of Child & Adolescent Psychiatry*, 51(3):238–248.
- Rutherford, M. D. & Towns, A. M. (2008). Scan path differences and similarities during emotion perception in those with and without autism spectrum disorders. *Journal of Autism and Developmental Disorders*, 38(7):1371–1381.
- Senju, A. & Johnson, M. H. (2009). Atypical eye contact in autism: models, mechanisms and development. *Neuroscience & Biobehavioral Reviews*, 33: 1204–1214.
- Simmons, D. R., Robertson, A. E., McKay, L. S., Toal, E., McAleer, P., & Pollick, F. E. (2009). Vision in autism spectrum disorders. *Vision Research*, 49(22): 2705–2739.
- Smilek, D., Birmingham, E., Cameron, D., Bischof, W., & Kingstone, A. (2006). Cognitive ethology and exploring attention in real-world scenes. *Brain Research*, 1080:101–119.
- Speer, L. L., Cook, A. E., McMahon, W. M., & Clark, E. (2007). Face processing in children with autism: Effects of stimulus contents and type. *Autism*, 11 (3):265–277.

## 1. Introduction

- Sterling, L., Dawson, G., Webb, S., Murias, M., Munson, J., Panagiotides, H., & Aylward, E. (2008). The role of face familiarity in eye tracking of faces by individuals with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, 38(9):1666–1675.
- Stone, W. L., Lee, E. B., Ashford, L., Brissie, J., Hepburn, S. L., Coonrod, E. E., & Weiss, B. H. (1999). Can autism be diagnosed accurately in children under 3 years? *Journal of Child Psychology and Psychiatry*, 40(2):219–226.
- Swinkels, S. H. N., Dietz, C., van Daalen, E., Kerkhof, I. H. G. M., van Engeland, H., & Buitelaar, J. K. (2006). Screening for autistic spectrum in children aged 14 to 15 months. I: The development of the early screening of autistic traits questionnaire (ESAT). *Journal of Autism and Developmental Disorders*, 36(6):723–732.
- Turner, L. M. & Stone, W. L. (2007). Variability in outcome for children with an ASD diagnosis at age 2. *Journal of Child Psychology and Psychiatry*, 48(8):793–802.
- van Berckelaer-Onnes, I. A., van de Blind, G., Anzion, P., & Werkgroep JGZ Richtlijn ASS. (2015). *JGZ-richtlijn Autismespectrumstoornissen. Signalering, begeleiding en toeleiding naar diagnostiek*. Trimboos-instituut.
- Van der Geest, J. N., Kemner, C., Verbaten, M. N., & van Engeland, H. (2002). Gaze behavior of children with pervasive developmental disorder toward human faces: a fixation time study. *Journal of Child Psychology and Psychiatry*, 43(5):669–678.
- Van Eylen, L., Boets, B., Steyaert, J., Wagemans, J., & Noens, I. (2015). Local and global visual processing in autism spectrum disorders: Influence of task and sample characteristics and relation to symptom severity. *Journal of Autism and Developmental Disorders*.
- Wass, S. V., Smith, T. J., & Johnson, M. H. (2013). Parsing eye-tracking data of variable quality to provide accurate fixation duration estimates in infants and adults. *Behavior Research Methods*, 45(1):229–250.
- Wiggins, L. D., Baio, J., & Rice, C. (2006). Examination of the time between first evaluation and first autism spectrum diagnosis in a population-based sample. *Developmental and Behavioral Pediatrics*, 27(2):S79–S87.
- Wolfe, J. M. (1998a). What can 1 million trials tell us about visual search? *Psychological Science*, 9(1):33–39.
- Wolfe, J. M. (1998b). Visual search. In Pashler, H., editor, *Attention*, pages 1–41. University College London Press, London, UK.

**Part I.**

**Advances in eye tracking  
methodology**



## **2. Qualitative tests of remote eye-tracker recovery and performance during head rotation**

Published as:

Hessels, R. S., Cornelissen, T. H. W., Kemner, C., & Hooge, I. T. C. (2015). Qualitative tests of remote eyetracker recovery and performance during head rotation. *Behavior Research Methods*, 47(3):848–859.

Author contributions:

RH, TC, CK, IH designed the study. RH, TC collected and analyzed the data. RH, TC, IH interpreted the data. RH drafted the paper. RH, TC, CK, IH finalized the paper.

## **Abstract**

What are the decision criteria for choosing an eye tracker? Often the choice is based on specifications by the manufacturer of the validity (accuracy) and reliability (precision) of measurements that can be achieved using a particular eye tracker. These specifications are mostly achieved under optimal conditions – for example, by using an artificial eye or trained participants fixed in a chin rest. Research, however, does not always take place in optimal conditions. For instance, when investigating eye movements in infants, school children, and patient groups with disorders such as attention-deficit hyperactivity disorder, it is practically impossible to restrict movement. We modeled movements often seen in infant research in two behaviors: 1) looking away from and back to the screen, to investigate eye tracker recovery, and 2) head orientations, to investigate eye tracker performance with non-optimal orientations of the eyes. We investigated how eight eye-tracking setups by three manufacturers (SMI, Tobii, and LC Technologies) coped with these modeled behaviors in adults. We report that the tested SMI eye trackers dropped in sampling frequency when the eyes were not visible to the eye tracker, whereas the other systems did not, and discuss the potential consequences thereof. Furthermore, we report that the tested eye trackers varied in their rates of data loss and systematic offsets during shifted head orientations. We conclude that (prospective) eye-movement researchers who cannot restrict movement or non-optimal head orientations in their participants might benefit from testing their eye tracker in non-optimal conditions. Additionally, researchers should be aware of the data loss and inaccuracies that might result from non-optimal head orientations.

Remote video-based eye trackers are growing in popularity among various research disciplines (Holmqvist et al., 2011), particularly because they are easy to set up and use. When choosing which remote eye tracker to use, researchers are faced with a plethora of options, all with slightly different technical specifications. Manufacturers specify how accurate their eye tracker is (spatial accuracy, the average offset between the point on screen the participant looks at and what the eye tracker reports), how reliable a measurement is (spatial precision, the sample to sample difference while the eye remains still), and in what range of distances to the eye tracker tracking of the eyes is possible (headbox dimensions). In addition, manufacturers constantly improve their eye trackers and aim for the best specifications possible. This alone leaves an individual researcher with choices of which the consequences might be difficult to grasp.

The specifications presented by the eye tracker manufacturers are, furthermore, often achieved under optimal conditions. Optimal conditions are, for instance, a fixed amount of light in the room, restricting a human participant from moving, or using an artificial eye instead of a human participant. Research, in contrast, does not take place in optimal conditions. In a recent attempt to find the most suitable eye tracker for a prospective infant study, we reached the conclusion that manufacturer specifications were not informative enough. We knew beforehand that our infant participants would not be measured in the manufacturers' optimal conditions. We realized that the problem of choosing between eye-tracker characteristics and measuring in non-optimal conditions goes beyond our infant research, but applies to a much broader range of participant groups.

The problem is best illustrated with an example. Let's consider a binocular measurement (i.e. tracking both eyes) of an infant participant using a Tobii TX300, a common eye tracker in infant research. According to the Tobii specifications, a spatial accuracy of  $0.4^\circ$  (which is an average offset of gaze position of 0.4 cm on screen at 57 cm viewing distance) and a spatial precision of  $0.09^\circ$  are achieved under a specific amount of light with a participant fixed in the center of the head box by means of a chin rest.

## 2. *Eye-tracker recovery and performance*

However, in a more realistic lab setting both non-optimal lighting conditions and an infant not fixed in a chinrest (i.e. an infant that is able to move), will deteriorate the accuracy of the eye tracker. Furthermore, sub-optimal calibration with the infant due to large calibration stimuli (which is common in infant research) affects eye-tracker accuracy even further, since it is impossible to know where exactly the infant is looking (e.g. at the top or bottom of the calibration stimulus) while calibrating. These examples apply not only to infant studies: In several research fields it is often not possible or even desirable to test a participant in optimal settings for various reasons, including ethical ones. This applies to any study in which the participant cannot be instructed to sit or be restrained in the optimal position in the eye trackers' head box, whether the participants are infants, school children, or patients with Down's syndrome, attention deficit hyperactivity disorder (ADHD), or muscular disorders. How then to interpret the technical specifications of an eye tracker for research in non-optimal circumstances? Can we assume that the eye tracker with the best specifications will still perform best when pushed beyond its comfort zone?

Here we propose a set of tests to qualitatively assess eye trackers' performance in non-optimal conditions, in order to aid potential users of eye trackers in their choice, and to indicate potential issues in interpretation of eye-tracking data. The focus is not on determining the best system, but specifically on whether eye trackers are robust to a set of head movements often seen in eye-tracking research with infants. As one of the optimal conditions for an eye tracker (as described above) is that a participant is positioned in the middle of the head box (i.e. the space in which reliable tracking of the eyes is possible), preferably moves as little as possible and looks straight at the screen, we were interested in eye trackers' performance during the position changes of the infant. We modeled the infants' changing head position in two behaviors: 1) looking away from the screen and back, and 2) shifting head orientations. During these movements and orientations we investigated whether the eye trackers still reported gaze data. If the system did report gaze data, we were interested in whether there

was any indication for systematic offsets (i.e. the same offset across trials and participants) or unsystematic offsets (i.e. highly variable across trials and participants). While these movements and orientations are inspired by infant research, they are relevant for any research field in which the participant cannot be instructed or positioned fully to the experimenters' liking. In addition to modeling these two behaviors, we investigate whether there are system-specific issues during these behaviors that are important for data analysis; for instance during the detection of periods in which the eye remains still (fixations) and periods of ballistic movement (saccades). We tested 8 different eye-tracking setups, and discuss our findings with regard to their applications to eye tracking in difficult groups such as infants, children, and certain patient groups.

## 2.1. Methods

### 2.1.1. Participants

A total of 9 volunteers participated in the study. Each of eight eye-tracking setups (see Table 2.1) was tested with 5 of these 9 volunteers. The setups were tested in two labs: at Utrecht University, the Netherlands and at Lund University, Sweden. Because of this, only 2 out of 9 participants (RH and TC; first and second author) could participate in all setups. Due to varying availability of the eye-tracking setups, the testing order was not identical for each participant. Mean age was 29.2 years ( $sd = 8.15$  years). All participants had normal or corrected-to-normal vision and reported no ocular deficits. 7 participants had previous experience with participating in and conducting eye-tracking research.

### 2.1.2. Apparatus

8 different eye-tracking setups from 3 different manufacturers (SMI, Tobii, & LC Technologies) were tested, all of which are in production as of this writing. This specific set was chosen for two reasons: 1) SMI and Tobii are two manufacturers of the most common eye trackers in North-West Europe and 2) these eye trackers are commonly used in the labs we are familiar with.

## 2. Eye-tracker recovery and performance

Table 2.1.: Participation of volunteers in each eye-tracking setup

Participant	Tobii X2-60	Tobii T120	Tobii TX300-120Hz	Tobii TX300-300Hz	SMI REDm-60Hz	SMI REDm-120Hz	SMI RED-250	LC Technologies EyeFollower
CF	✓		✓	✓	✓	✓	✓	
RH	✓	✓	✓	✓	✓	✓	✓	✓
TC	✓	✓	✓	✓	✓	✓	✓	✓
JL	✓		✓	✓	✓	✓		
LW	✓		✓	✓	✓	✓		
MN		✓					✓	✓
IH		✓					✓	✓
DW		✓					✓	
AS								✓

The LC Technologies EyeFollower was included as it is specifically designed to allow a large range of movement. No conflicts of interest with any of the manufacturers were present. Eye-tracker specifications as provided by the manufacturers are summarized in Table 2.2. While these specifications give a good overview of the general differences between the devices, we did not make any assumptions about an eye tracker’s performance based on them.

The Tobii X2-60 and SMI REDm eye trackers were the only eye trackers not attached to a screen. They were positioned at the bottom of a laptop display placed on a table, as they are most commonly used. As a result, participants looked slightly down at the screen with the Tobii X2-60 and SMI REDm compared to the other setups. The other eye trackers were integrated in a monitor, which was positioned perpendicular to the table, with the middle of the screen roughly at eye height.

Stimulus presentation was done using MATLAB and the PsychToolbox (Brainard, 1997). Data recording was done using the iView SDK for SMI, Tobii SDK for Tobii, and EyeGaze for LC Technologies. Hereafter, all data files were imported into MATLAB for data analysis.

### 2.1.3. Procedure

Participants were positioned in front of the eye tracker (see apparatus for more info on the different setups) at the optimal tracking distance for each setup. This was either done by using the distance values reported back by the eye-tracking software (i.e. Tobii SDK for Tobii, iView SDK for SMI), or by the experimenter positioning the participant at the optimal tracking distance reported by the manufacturer (for the LC Technologies EyeFollower). Hereafter, a calibration sequence was run. For the SMI systems a 5-point calibration was performed followed by a validation sequence. For the Tobii systems a 5-point calibration was run followed by an inspection of calibration results using the Tobii SDK. For the LC Technologies a 9-point calibration was run followed by a validation sequence. Calibration was repeated until quality of calibration was judged to be good enough

## 2. Eye-tracker recovery and performance

Table 2.2.: Specifications of each eye-tracking setup. Specifications are retrieved from the eye-tracker manufacturers' product descriptions, and based on binocular data, not processed or filtered after recording.

	Tobii X2-60	Tobii T120	Tobii TX300-120Hz	Tobii TX300-300Hz	SMI REDm-60Hz	SMI REDm-120Hz	SMI RED 250	LC Technologies EyeFollower
Sampling rate	60 Hz	120 Hz	120 Hz	300 Hz	60 Hz	120 Hz	250 Hz	120 Hz interlaced
Accuracy	0.4°	0.4°	0.4°	0.4°	0.5°	0.5°	0.4°	0.4°
Precision	0.34°	0.16°	0.07°	0.07°	0.1°	0.1°	0.03°	1
Headbox	50x36 cm (70 cm)	30x22 cm (70 cm)	37x17 cm (65 cm)	37x17 cm (65 cm)	32x21 cm (60 cm)	32x21 cm (60 cm)	40x20 cm (70 cm)	76x50 cm
Tracking distance	40-90 cm	50-80 cm	50-80 cm	50-80 cm	50-75 cm	50-75 cm	60-80 cm	46-97 cm
Price <sup>2</sup>	15,500 €	25,900 €	32,900 €	32,900 €	17,500 €	17,500 €	25,900 €	24,400 €

<sup>1</sup> Not specified by manufacturer

<sup>2</sup> Price as quoted to us / our department. Excludes both VAT and any discounts.

by the experimenters. After positioning and calibration, participants were presented with three tasks: a ‘recovery’ task, a ‘yaw orientation’ task, and a ‘roll orientation’ task.

The main interest for the recovery task was to determine what happens when an eye tracker loses track of the eyes (i.e. theoretically can not report gaze data anymore) and when it restarts reporting gaze data. The focus here is on how an eye tracker recovers not when. The main interest for the yaw orientation task and the roll orientation task was to determine how eye trackers cope with eyes in non-optimal head orientations.

### **Recovery task**

Each trial consisted of a 5 second period, during which a fixation dot was presented in the center of the screen. Participants were instructed to look at the fixation dot whenever they looked at the screen. After 1 second a low-pitched beep sounded, and after another 2 seconds a high-pitched beep sounded. Prior to starting the task, participants were instructed to turn their head to the left or right at the low beep, and to turn back to their starting position at the high beep. After fixating for another 2 seconds the next trial followed.

### **Yaw orientation task**

Each trial again consisted of a 5 second period with a low-pitched beep after 1 second, and a high-pitched beep after another 2 seconds. Prior to starting the task, participants were instructed to turn their head to the left or to the right as far as possible while maintaining fixation on the fixation dot, and to turn back to their starting orientation at the high beep. Figure 2.1 depicts the axis along which the head rotation takes place.

### **Roll orientation task**

Each trial again consisted of a 5 second period with a low-pitched beep after 1 second, and a high-pitched beep after another 2 seconds. Prior

## 2. Eye-tracker recovery and performance

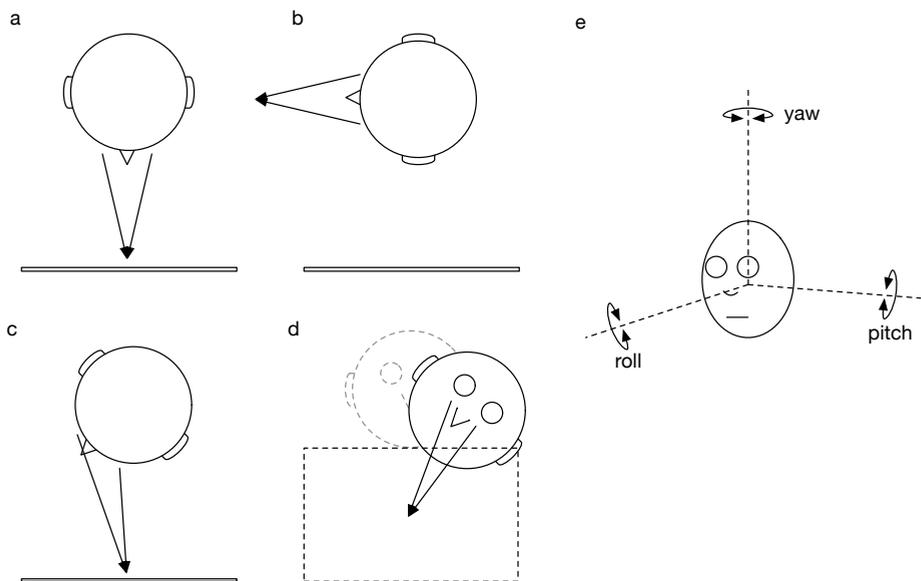


Figure 2.1.: Schematic overview of head positions. (a) Top view of starting and ending positions. (b & c) Top views of positions between beeps in (b) the recovery task and (c) the yaw orientation task. (d) Front view of a position between beeps in the roll orientation task. The dotted rectangle in this panel represents the screen and is moved down relative to its original position in the experiment for clarification purposes only. (e) Axes along which the head can rotate. Rotations along the yaw and roll axes are included in the present study in the yaw orientation and roll orientation tasks, respectively.

to starting the task, participants were instructed to tilt their head to the left or to the right while maintaining fixation on the fixation dot at the low beep, and to turn back to their starting orientation at the high beep. Figure 2.1 depicts the axis along which the head rotation takes place.

Participants were asked to hold fixation on a central point on screen so that we could compare offset of gaze data to this central point prior to movement and after movement; where we would expect offsets to be minimal at least prior to movement. Schematic overviews of final head positions in all tasks, as well as the axes along which head rotations took place, are given in Figure 2.1. For each task, participants started with 10 trials with head movements to the left, followed by 10 trials to the right (by instruction of the experimenter). To minimize the amount of movements in the wrong direction made by participants, the order of direction and the order of tasks, were not counterbalanced. In addition, the order of movement direction was not counterbalanced across participants. This meant that the order of movement direction was identical for all participants in each eye-tracking setup. No instructions were given to the participants with regard to blinking.

While the head movements and orientations in the tasks presented here are far from ideal for the eye trackers, we did measure with highly motivated, cooperating participants who understood the instructions and who were all familiar with psychological research using eye trackers.

## 2.2. Results

The results section is divided into four sections: three sections with point of regard (i.e. the gaze position reported by the eye tracker) from the three separate tasks and one section on data loss. We consider first, however, the results that we would expect if an eye tracker performs perfectly in all tasks. In the recovery task we would expect point of regard to start on the fixation dot. Hereafter, we would expect to see the point of regard moving to the left or right as gaze moves off screen. Finally, we would

## 2. *Eye-tracker recovery and performance*

expect the reverse once the participant returns gaze back to the screen. For the yaw orientation task and the roll orientation task we would expect point of regard to be on the fixation dot throughout the trial, as no eye movements away from the fixation dot are made. We, would, however, expect some minor changes in point of regard during the head movements themselves, as the eyes will have to correct their orientation for the change of head orientation. After the results from the recovery, yaw orientation, and roll orientation tasks are discussed, data loss in the three tasks will be presented: this to substantiate the qualitative results we outline in the first three sections and establish the robustness of eye trackers' gaze reporting in the two orientation tasks.

### 2.2.1. **Recovery task**

The main purpose for the recovery task was to determine what happens when eye trackers lose track of and regain tracking of the eyes. Figure 2.2 depicts the horizontal screen coordinates as reported by the eye tracker, for all trials from the recovery task, separated for all eye-tracker setups. Only samples of which the coordinates were on screen are depicted in Figure 2.2; the vertical axis depicts the entire screen width. Between 0 and 1 second after trial onset participants remained fixated on the middle of the screen, indicated by the coordinates being in the center of the vertical axis. When the first beep sounded (i.e. the first vertical bar in the graph), participants looked away from the screen. Between 2 and 3 seconds after trial onset the maximum position of the head orientation was reached in almost all the trials (i.e. the gaze is completely turned away from the screen and no point of regard is reported by the eye tracker). In nearly all the setups, the eye tracker was capable of following the gaze off-screen, as is visible from the gaze coordinates moving from center to the edge of the screen (which is also the edge of the graph). The eye trackers with higher sample frequency (120 Hz and higher) obviously collect more samples from onset to offset of movement, which means detection of looking away is easier in these systems: there is more information available to do so because there are more samples during the movement. When participants returned

to fixation after 3 seconds (indicated by the second vertical bar), there were several remarkable differences between setups. Only three eye-tracker setups were able to track the gaze of the participant immediately upon gaze reentering the screen as visible from the horizontal coordinate moving from the edge of the graph back to the center: the Tobii TX300 (both at 120 Hz and at 300 Hz), and the LC Technologies EyeFollower. If mere sampling frequency were the determining factor in being able to track the eyes immediately when the eyes return to the screen, one would expect the SMI REDm at 120 Hz, the Tobii T120, and the SMI RED250 to be able to do this as well. In these latter systems however, it was impossible to detect the gaze returning back to the fixation point on screen before the point of regard hit the fixation dot. This led us to question what the difference was for the eye tracker prior to and after the gaze shift; the gaze shift is identical, only in the opposite direction.

As the SMI RED250 and Tobii TX300 have similar sampling frequencies, 250 Hz versus 300 Hz, we expected similar behavior from the eye tracker in terms of tracking the eyes upon returning to the screen. As the SMI RED250 and Tobii TX300 showed different behavior, we decided to take a closer look at what these systems report when a participant looks away. Two typical trials in the recovery task from these two systems are depicted in Figure 2.3. As visible in the top graph, the SMI RED250 showed an increase in inter-sample interval shortly ( $\approx 500\text{ms}$ ) after the gaze was shifted away from the screen completely. This means that during this period the sampling frequency dropped sharply from the manufacturer-specified 250 Hz and stabilizes at 20 Hz while the gaze remained turned away. When the gaze returned to the screen, the inter-sample interval spiked, after which it returned to 250 Hz and gaze data was again reported. In addition to the SMI RED250, the SMI REDm (both at 60 Hz and at 120 Hz) showed the same pattern. The Tobii TX300, on the other hand, continued to report empty samples at 300 Hz when the eyes were lost, and reported gaze data shortly after the gaze was back on the screen. While the reasons for this difference are technical in nature and beyond the scope of this paper, whether an eye tracker drops in sampling frequency or not when movements

## 2. Eye-tracker recovery and performance

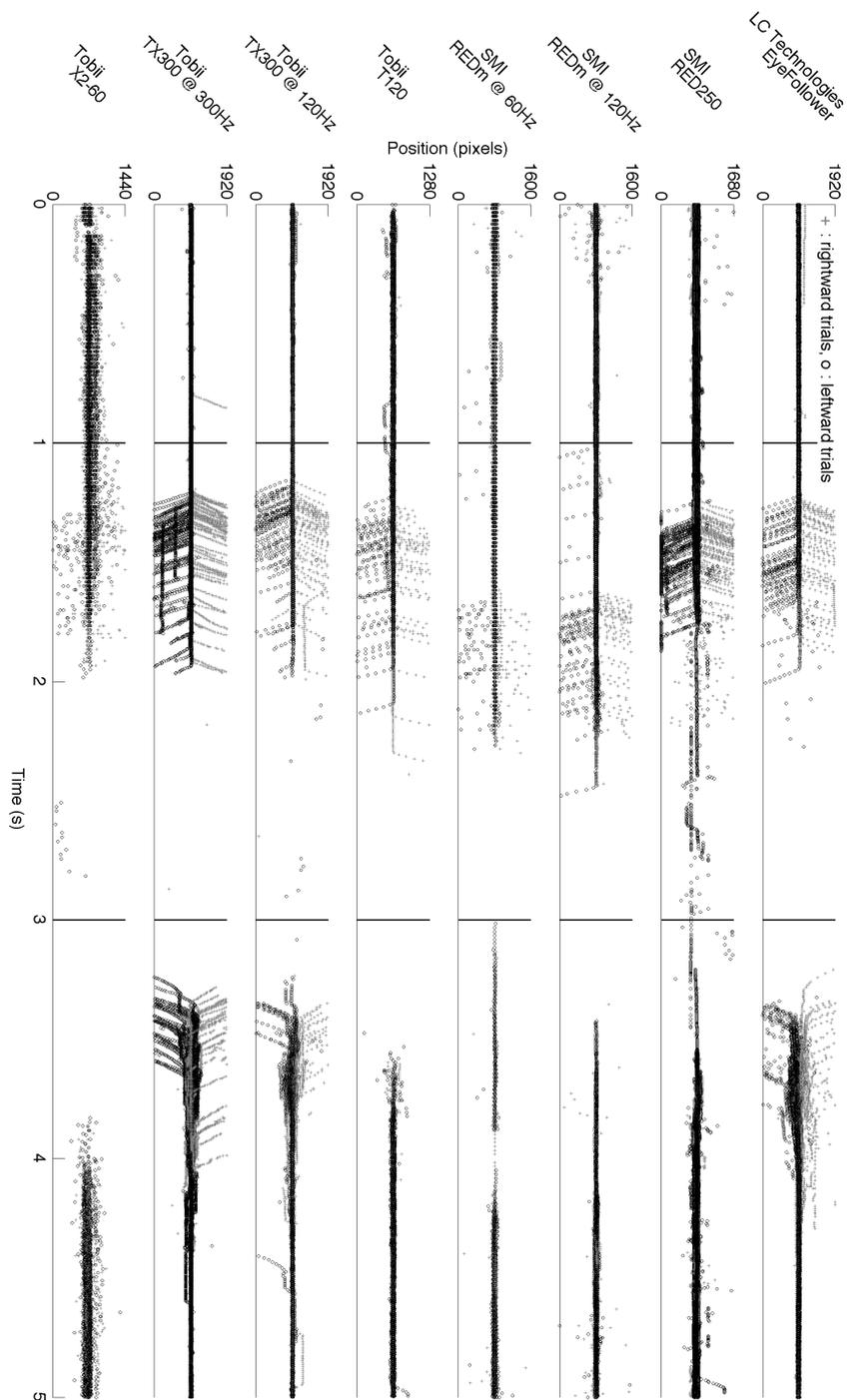


Figure 2.2.: Data from the recovery task for each eye-tracker setup. Raw data from both eyes were averaged for all setups except the LC Technologies EyeFollower, which gives alternating coordinates from left and right eyes every other sample. All trials from five participants are overlaid. Since the movement in Task 1 was only in the horizontal direction, only horizontal coordinates are given. Black circles indicate trials with rightward movement, and gray crosses indicate trials with leftward movement. Black vertical bars indicate the beeps that signaled a movement to be made away from the screen at 1 s, and back to the screen at 3 s. Due to the audio latency in the SMI REDm/MATLAB setup, the trials here are shifted rightward slightly.

## 2. Eye-tracker recovery and performance

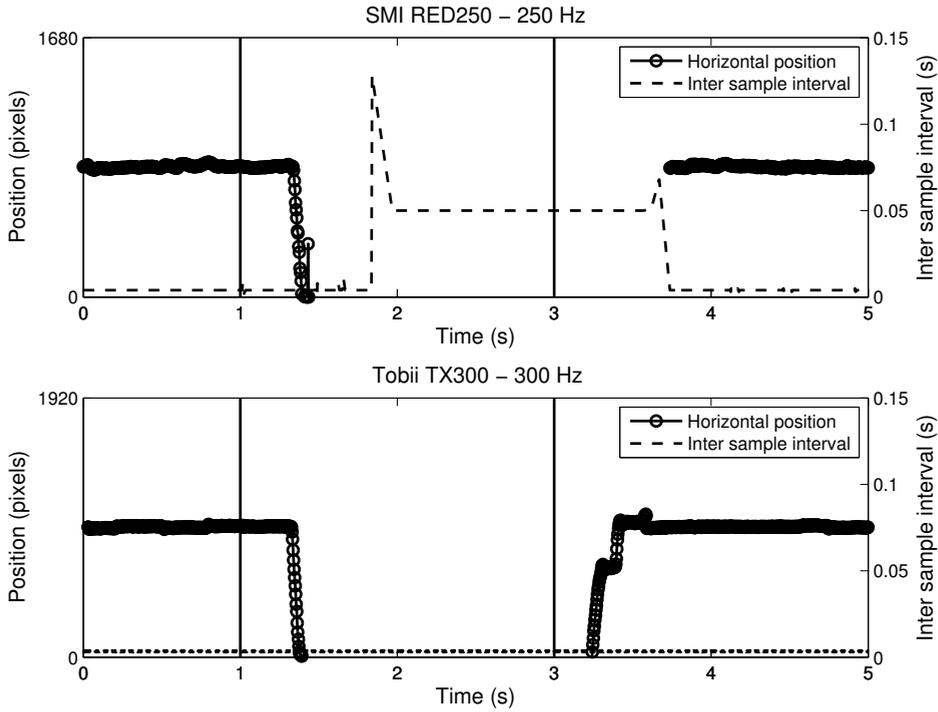


Figure 2.3.: Horizontal positions and inter-sample intervals in two typical trials from the recovery task in the SMI RED250 and the Tobii TX300 at 300 Hz. In the SMI RED250 system, the inter-sample interval increases once the head is shifted completely away from the screen, whereas this does not occur in the Tobii TX300.

are made might be important for both the detection of fixations, saccades, and smooth pursuit (event detection). In addition, it might be important for detecting gaze shifts back to the screen.

### 2.2.2. Yaw orientation

The main purpose for the yaw orientation task was to determine how the eye trackers dealt with non-optimal head orientations; in this case the head turned sideways, with the gaze still on the screen. In this orientation, one of the eyes moves (partially) behind the nasion. The primary question was

whether the eye tracker was able to report gaze data during the non-optimal orientation, and whether there was any indication of (un)systematic errors when doing so. Figure 2.4 depicts the horizontal screen coordinates for all trials from the yaw orientation task for participants RH and TC in all eye-tracker setups. Only the two subjects that participated in all 8 setups are pictured, as we wanted to determine whether offset between the point of regard as reported by the eye tracker and the fixation dot were similar or different across participants. While only data from participants RH and TC are described, the pattern of eye-tracker performance was comparable across all 5 participants in each setup.

Particularly the SMI RED250 and the Tobii T120 struggled during the yaw orientation task: as seen in Figure 2.4 both trackers appeared to report little data between 2 and 3 seconds, during which the difference in head orientation from the starting orientation was maximal. The LC Technologies EyeFollower, SMI REDm, Tobii TX300, and Tobii X2-60 were able to calculate point of regard, albeit with an offset from the fixation dot. The point of regard signal from the SMI REDm, both at 60 Hz and 120 Hz, showed large offsets from the fixation dot during the shifted head orientation. Surprisingly, this shift seemed very persistent in some trials, causing large offsets in point of regard even when participants moved their head back to the starting orientation. This is visible from Figure 2.4, where the SMI REDm at 120 Hz for participant TC showed a persistent offset in point of regard (i.e. the second black line shifted upwards from the line in the center of the screen). This occurred even between 0 and 1 seconds when the participant was not positioned in a non-optimal head orientation. This occurred for participant TC, but not for participant RH, yet also for other participants. These large offsets likely resulted from the SMI REDm system switching the position of the eyes. If the head is rotated left, the position of the right eye moves towards where the left eye was. If the left eye is then lost from the eye image, the right eye is mistaken for the left, hence a large but systematic shift in point of regard occurs. This seemed to happen with the SMI REDm on several occasions, and the switch appeared

## 2. Eye-tracker recovery and performance

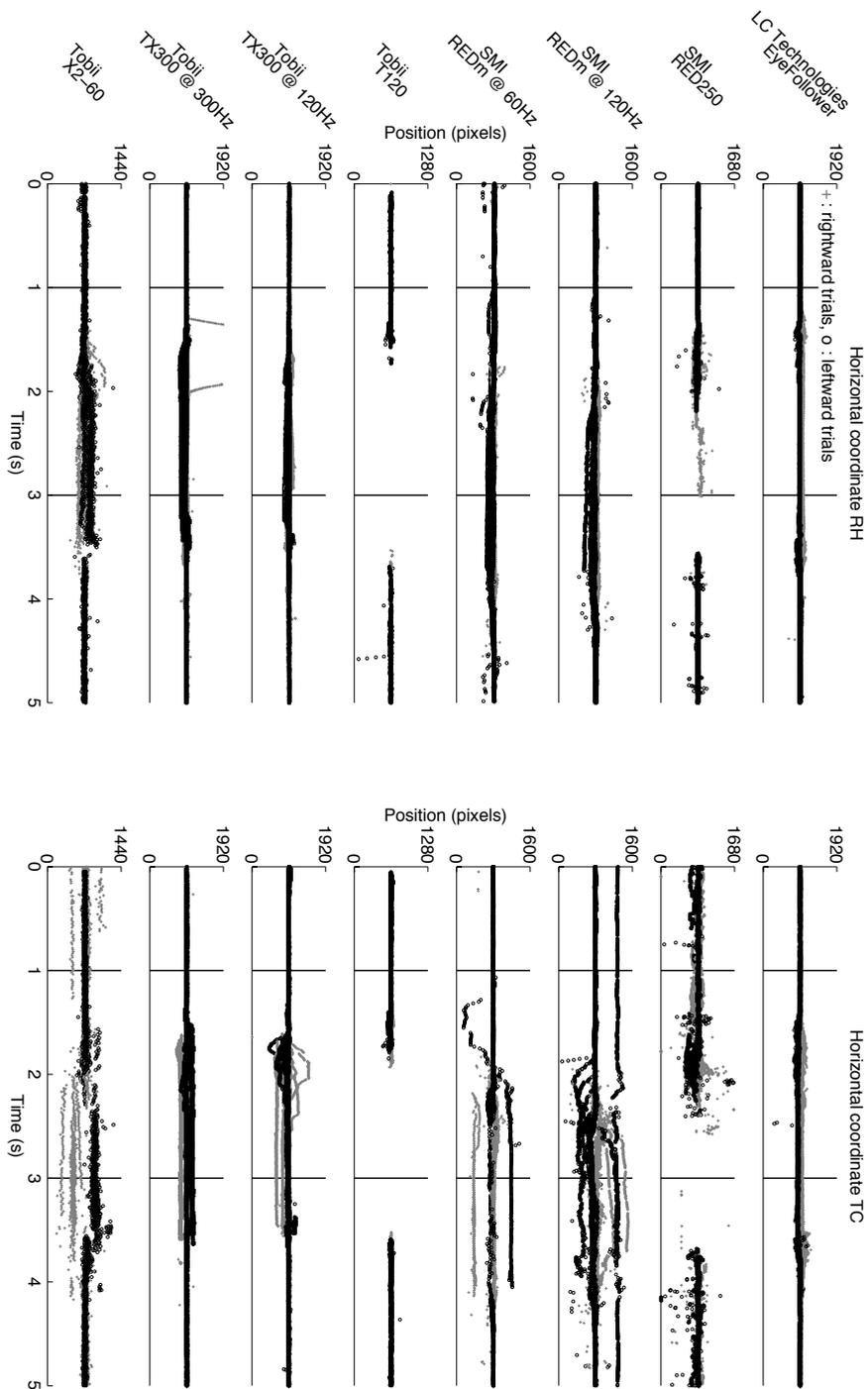


Figure 2.4.: Data from the yaw orientation task for each eye-tracker setup. Raw data from both eyes were averaged for all setups except the LC Technologies EyeFollower, which gives alternating coordinates from the left and right eyes every other sample. All trials from observer R.H. (left) and observer T.C. (right) are overlaid. As in Task 1, only horizontal coordinates are given, because the horizontal coordinates showed the most interesting results. Black circles indicate trials with rightward rotation, and gray crosses indicate trials with leftward rotation. Black vertical bars indicate the beeps that signaled a rotation to be made away from the screen at 1 s and back to the screen at 3 s. Due to the audio latency in the SMI REDm/MATLAB setup, the trials here are shifted rightward slightly.

## 2. *Eye-tracker recovery and performance*

quite persistent after point of regard returned to the starting position<sup>3</sup>.

### 2.2.3. Roll orientation task

The main purpose for the roll orientation task was identical to the purpose of the yaw orientation task: to determine how eye trackers dealt with non-optimal head orientations. The primary question was whether the eye tracker was able to report gaze data during the non-optimal orientation, and whether there was any indication of (un)systematic errors when doing so. Figure 2.5 depicts the horizontal screen coordinates for all trials from the roll orientation task for participants RH and TC in all eye-tracker setups. As in the yaw orientation task, only the two participants who participated in all 8 setups are pictured, and only horizontal coordinates are given.

As in the yaw orientation task, the Tobii T120 appeared to struggle with the head orientations in the roll orientation task, losing almost all data after movement in participant RH and TC. The SMI RED250 also appeared to struggle, losing a lot of data during the non-optimal head orientation especially with participant TC, though less than in the yaw orientation task. The eye-tracker systems that did continue to report gaze data during the shifted head orientation appeared to report larger offsets from the fixation dot than in the yaw orientation task. Particularly the LC Technologies EyeFollower reported large offsets, although they seemed to be very systematic for participant RH: the offsets from the fixation dot were consistent in direction and amplitude over trials.

### 2.2.4. Data loss

In addition to the qualitative results described above, we attempted to give some quantification to the performance of the eye trackers in our tasks and calculated the proportion of data loss over all participants in two time periods. Data loss was defined as the number of samples in a time period in

---

<sup>3</sup>SMI is aware of this issue and is working on it (Pötter 2014, personal communication)

which point of regard (i.e. the gaze position reported by the eye tracker) was not reported for either eye, divided by the theoretical number of samples in that time period. This number would, for example, be 250 samples for the RED250 in one second, although this might not be the actual number of samples that was recorded due to a drop in sampling rate. The first time period is between 0 and 1 second after trial onset. In this period participants fixated on the screen center, and no movement occurred. We would expect little to no data loss here. The second period is between 2 and 3 seconds after trial onset. In this period, head movement occurred, and the difference to the first time period was expected to be largest. We would expect the data loss to be highest here if an eye tracker cannot deal with a specific movement, and to be 0 if an eye tracker can deal with the specific movement. There was no gaze on screen in the recovery task between 2 and 3 seconds, and we included data loss in this task as a comparison: We expect little to no data loss in the period 0-1 seconds, and near maximum data loss in the period 2-3 seconds. Data loss for all three tasks is given in Figure 2.6.

In the recovery task, the proportion of data loss during 2-3 seconds was near 1 for all eye-tracker setups, although slightly lower for the SMI REDm. However, this could be explained by the SMI REDm / MATLAB setup including a slightly longer audio latency, meaning that gaze shifts might have started later. Between 0-1 seconds, proportion of data loss was between 0 and 0.2 for all eye-tracker setups. In addition, our qualitative assessment in the yaw orientation task that the SMI RED250 and Tobii T120 were struggling most, was confirmed by the high proportion data loss (between 0.6 and 0.85) reported in Figure 2.6. The LC Technologies EyeFollower reported a high proportion of data loss despite reporting fairly stable gaze data. As reported before, in the yaw orientation task one of the eyes moves (partially) behind the nasion. The LC Technologies EyeFollower measures at 120 Hz interlaced (i.e. sample from each eye every other sample), and as one eye was behind the nasion, we only obtained valid samples every other sample (i.e. from the eye that was still visible). We therefore expected to obtain a proportion of data loss nearing 0.5, which is confirmed in Figure

## 2. Eye-tracker recovery and performance

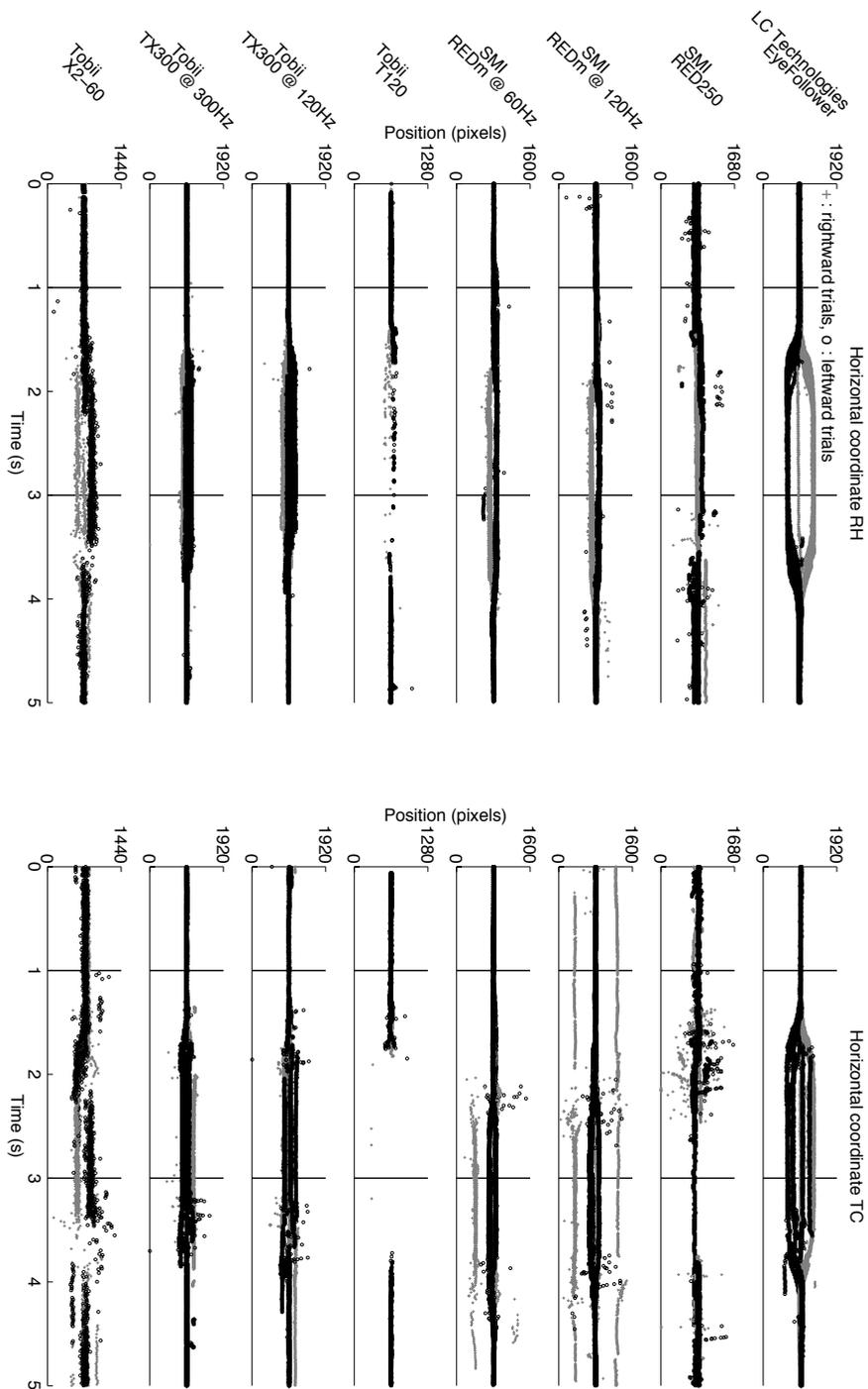


Figure 2.5.: Data from the roll orientation task for each eye-tracker setup. Raw data from both eyes were averaged for all setups except the LC Technologies EyeFollower, which gives alternating coordinates from the left and right eyes every other sample. All trials from observer R.H. (left) and observer T.C. (right) are overlaid. As in Tasks 1 and 2, only horizontal coordinates are given, because the horizontal coordinates showed the most interesting results. Black circles indicate trials with rightward rotation, and gray crosses indicate trials with leftward rotation. Black vertical bars indicate the beeps that signaled a rotation to be made away from the screen at 1 s and back to the screen at 3 s. Due to the audio latency in the SMI REDm/MATLAB setup, the trials here are shifted rightward slightly.

## 2. *Eye-tracker recovery and performance*

2.6. In the roll orientation task, the SMI RED250 and the Tobii T120 again reported the highest proportion of data loss (between 0.5 and 0.9). The proportion of data loss for the LC Technologies EyeFollower was low in the roll orientation task, as compared to in yaw orientation task. In the roll orientation task, both eyes remained directly visible to the eye-tracker camera, provided that the trackers' headbox was large enough. It is therefore not surprising that, unlike in the yaw orientation task where one eye moves behind the nasion, the proportion of data loss for the LC Technologies EyeFollower was low. The proportions of data loss for the SMI REDm setups, Tobii TX300 setups, and Tobii X2-60 were similar across the two orientation tasks.

## 2.3. Discussion

The aim of the present study was to provide a set of qualitative tests for judging eye-tracker performance when head movements and rotations are made. These movements and orientations were inspired by infant eye-tracking research: infants tend to be distracted easily, and often make gaze shifts away from and back to the screen. Additionally, it is practically impossible to restrain infants' movements in front of the eye tracker, which results in non-optimal head orientations (i.e. a rotation other than looking straight ahead at the eye tracker). We modeled movements often seen in infant research in two behaviors 1) looking away from the screen and back to it, and 2) head rotations, in order to determine how eye trackers cope with loss of tracking the eyes in non-optimal head positions. The head movements performed in the present tests are, however, not solely relevant to infant research. Any eye-tracking research field in which participants cannot be fully instructed or positioned to the researchers' liking will encounter non-optimal head orientations: For example when studying school children, patients with Down's syndrome, ADHD, or muscular disorders.

We report that it was possible with most eye trackers to detect a gaze shift away from the screen, although it might be done more reliably when sampling frequency was high (i.e. 120 Hz or higher), as a result of more

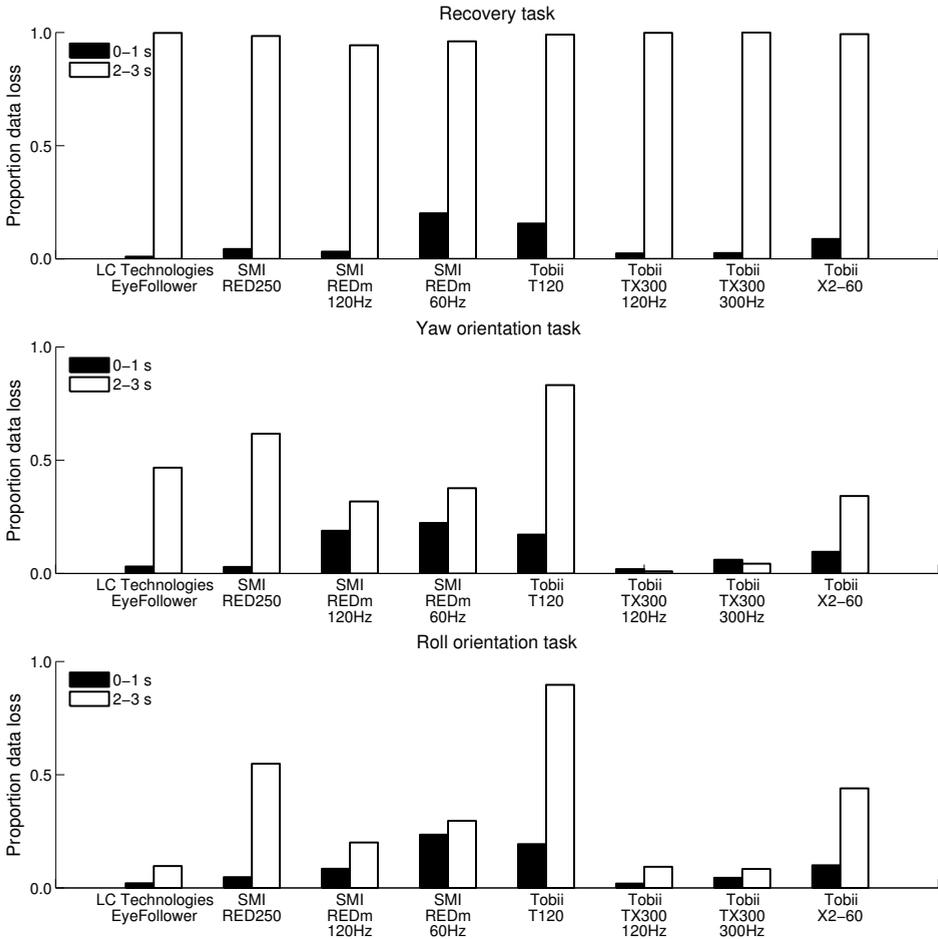


Figure 2.6.: Proportions of data loss for all eye-tracker setups in the recovery, yaw orientation, and roll orientation tasks. Data loss was calculated in two periods: 0-1 s, when data loss was expected to be minimal, and 2-3 s, when data loss was expected to be maximal in a trial. Data loss was calculated by dividing the number of samples in which at least one eye was found by the theoretical number of samples we would expect in a second on the basis of the sampling frequency. For example, for the SMI RED250 we expected 250 samples in 1 s; see also the comparison of the SMI RED250 versus the Tobii TX300 in a trial in Figure 2.3

## 2. *Eye-tracker recovery and performance*

samples, and thus more information, being available during the shift. In our set of eye-tracking setups, detecting a gaze shift back to the screen based on horizontal gaze coordinates during the shift could be done with both the LC Technologies EyeFollower and Tobii TX300. This is potentially useful if one is interested in showing a dynamic stimulus for a fixed viewing time. If one wants to pause it once an infant looks away and un-pause once the infant looks back, the gaze shift information has to be available. Implementing this could be done, for instance, by building a detector that is able to identify the gaze shift to the screen as it returns in the Tobii TX300 trial in Figure 2.3. This, as opposed to the gaze coordinate after returning to the screen in the SMI RED250 trial in Figure 2.3, where no point of regard is reported during the gaze shift back to the screen. The latter could occur also as a result of technical difficulties during the measurement: i.e. a participant who is actually looking continuously at the screen, but the eye tracker not being able to report point of regard.

We furthermore report that SMI eye trackers dropped in sampling frequency roughly 500 ms after the eye tracker cannot find the eyes and does not report gaze data anymore. The Tobii eye trackers, on the other hand, did not drop in sampling frequency. We can speculate on two different situations in which the eyes are lost. If the eyes are lost due to a blink or a hand moving in front of the eyes for less than 500 ms, we would expect the SMI eye trackers to recover as quickly as the Tobii eye trackers. If the eyes are lost due to a gaze shift away from the screen, which in infants are typically longer than 500 ms, we would expect the SMI eye trackers to recover more slowly than the Tobii eye trackers. Whether an eye tracker drops in sampling frequency might also be important to take into account for two other forms of data analysis. First, when detecting periods of movements and stillness of the eyes (often referred to as event detection) if drops in sampling frequency occur also when gaze samples are reported. Second, when one interpolates over periods of data loss. Interpolating is often done in infant research (see e.g. Frank, Vul, & Johnson, 2009; Saez de Urabain, Johnson, & Smith, 2015 for different methods of interpolating data loss), where short periods of data loss are more common than in adult studies.

In event detection, one may categorize periods of stillness of the eyes as fixations, periods of ballistic movement as saccades, and periods of constant movement as smooth pursuit (for instance when the eyes are following a moving object). Event detection is often done by calculating the velocity of the eye across samples (Holmqvist et al., 2011): by taking the distance between samples on screen and dividing it by the inter-sample interval. If one assumes that the inter-sample interval is fixed (i.e. 4ms in the case of a 250 Hz SMI RED250), velocity values, and consequently the decision whether a sample belongs to a fixation or saccade, will be different from when the actual inter-sample intervals as reported in the eye-tracker output are used. Changes in sampling frequency are thus vital for valid detection of events. As a consequence of not taking into account drops in sampling frequency, all metrics based on the timing of the detected events might be invalid too; for instance when one calculates the total time spent looking at an area of interest.

Eye trackers vary in how robust (i.e. whether they report gaze data, and whether it is accurate) they are to non-optimal head orientations. Both the SMI RED250 and Tobii T120 lost a lot of data during non-optimal head orientations, whereas the other eye-tracker setups did report gaze data. Particularly the Tobii TX300 reported very little data loss in both the yaw orientation task and the roll orientation task. Even when an eye tracker was able to calculate point of regard in a non-optimal head orientation, there appeared to be large offsets compared to optimal head orientation. These offsets were quite systematic for some systems (i.e. in the roll orientation task for the LC Technologies EyeFollower), and less systematic for others. As we have no reason to assume that participants were in fact looking somewhere else on screen but the fixation dot, we therefore assume that this offset is a result of reduced accuracy of the eye tracker. This would mean that although an eye tracker is robust to a certain movement and continues to report gaze data, accuracy might very well deteriorate.

If a participant is in a non-optimal orientation for the eye tracker and accuracy drops, this should be taken into account when creating areas of

## 2. *Eye-tracker recovery and performance*

interests for eye-tracking data analysis. When doing area of interest analysis, even a change in accuracy of  $0.5^\circ$  can significantly alter the outcomes (Holmqvist, Nyström, & Mulvey, 2012). If one disregards the shifted orientations of one's participants, and assumes data is accurate, study outcomes might be invalid. Consider, for example, a study assessing the time participants spent looking at the nose of a static face that is presented in the middle of the screen. If the participant is orientated non-optimally, accuracy might deteriorate and the time spent looking at the nose might be shifted up or down towards the eyes or mouth. This would result in a lower total time spent looking at the nose, than would have been observed when the participant was positioned optimally. One possible solution is to increase the size of areas of interest, although this depends on the stimulus one uses (see e.g. Holmqvist et al., 2011 for a more detailed discussion on creating areas of interest). One could furthermore discuss whether reporting highly inaccurate data is better than reporting no data at all when the participant is in a non-optimal head orientation. Regardless of whether it is preferable to obtain inaccurate data during non-optimal head orientations or not, researchers should be aware that the head orientation of their participant could be a factor in accuracy. The present study should be helpful in determining whether this is something to be wary of in one's specific situation or not.

The present study provides a first assessment of several eye trackers in non-optimal conditions. There are, however, several limitations that should be noted. First, the movements in the three tasks were performed across two different axes of movement, but independent measures of head rotation across these axes were lacking. Future research might benefit from more controlled measurements of rotation in each direction, to better investigate eye-tracker performance during non-optimal head rotations. Secondly, no specific instructions with regards to blinking were given, nor were blinks removed from the analyses. While blinks might have contributed slightly to measures of data loss, we reason that the amount of data loss due to one or more blinks is negligible compared to the periods of data loss we observed.

As research does not always take place in optimal conditions, eye-tracker accuracy, precision and sampling frequency might not be the best guide when deciding which eye tracker to use. We have proposed here a number of qualitative tests to investigate eye-tracker recovery and robustness to non-optimal head orientations, and highlighted several eye-tracker-specific issues. These tests, combined with eye-tracker specifications should provide a good overview of what an eye tracker can and cannot do, and could help make a more informed choice when choosing an eye tracker. Furthermore, they could provide ideas to test one's own eye tracker for suitability in doing research where the movements we highlighted may occur. We conclude that while eye-tracker specifications are certainly important, they are not necessarily decisive factors when one knows the research will not take place in optimal conditions. Eye-tracker performance in non-optimal conditions can be just as important.

## **Acknowledgements**

We would like to thank Kenneth Holmqvist, Fiona Mulvey, and the Eye-Tracking Group at the Lund University Humanities Lab for providing eye-tracking setups in this study. We would also like to thank Edwin Dalmaijer for help with the manuscript. This work was supported by a Netherlands Organization for Scientific Research (NWO) VICI grant (45307004) and an NWO Gravitation Grant (024.001.003), both to Chantal Kemner.

## References

- Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, 10(4):433–436.
- Frank, M. C., Vul, E., & Johnson, S. P. (2009). Development of infants' attention to faces during the first year. *Cognition*, 110(2):160–170.
- Holmqvist, K., Nyström, M., & Mulvey, F. (2012). Eye tracker data quality: What it is and how to measure it. *Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '12*, 45.
- Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press.
- Saez de Urabain, I. R., Johnson, M. H., & Smith, T. J. (2015). GraFIX: A semiautomatic approach for parsing low- and high-quality eye-tracking data. *Behavior Research Methods*, 47(1):53–72.

### **3. Consequences of eye color, positioning, and head movement for eye-tracking data quality in infant research**

Published as:

Hessels, R. S., Andersson, R., Hooge, I. T. C., Nyström, M., & Kemner, C. (2015). Consequences of eye color, positioning, and head movement for eye-tracking data quality in infant research. *Infancy*, 20(6):601–633.

Author contributions:

RH, RA, IH, MN, CK designed the study. Data was collected by research assistants at the KKC under supervision of RH. RH, RA analyzed the data. RH, RA, IH, MN, CK interpreted the data. RH drafted the paper. RH, RA, IH, MN, CK finalized the paper.

## **Abstract**

Eye tracking has become a valuable tool for investigating infant looking behavior over the last decades. However, while eye-tracking methodology and achieving high data quality have received much attention for adult participants, it is unclear how these results generalize to infant research. This is particularly important as infants behave different from adults in front of the eye tracker. In this study, we investigated whether eye physiology, positioning, and infant behavior affect measures of eye-tracking data quality: accuracy, precision, and data loss. We report that accuracy and precision are lower, and more data loss occurs for infants with bluish eye color compared to infants with brownish eye color. Moreover, accuracy was lower for infants positioned in a high chair or in the parents' lap compared to infants positioned in a baby seat. Finally, precision decreased and data loss increased as a function of time. We highlight the importance of data quality when comparing multiple groups, as differences in data quality can affect eye-tracking measures. In addition, we investigated how two different measures to quantify infant movement influence eye-tracker data quality. These findings might help researchers with data collection and help manufacturers develop better eye-tracking systems for infants.

According to Aslin (2007), *“It is no exaggeration to say that without looking time measures, we would know very little about infant development”* (p. 48). Over the years, researchers have employed measures of looking behavior to investigate infants’ detection, discrimination, and preferences for visual stimuli in a time when they cannot express their preferences or respond verbally (Aslin, 2007). In earlier studies observers would code where infants were looking to estimate global looking times at visual stimuli. The rise of eye tracking as a tool in infant research has induced a shift from investigating the macrostructure to the microstructure of infant looking behavior (Aslin, 2007; 2012). Eye trackers are now used to study, for example, infant oculomotor characteristics (Wass & Smith, 2014), object perception (Amso & Johnson, 2006), face processing in typically developing infants (Wheeler et al., 2011), and in infants at risk for autism spectrum disorders (Jones & Klin, 2013).

Remote video-based eye trackers are devices that commonly illuminate the eye with an infrared light that reflects off the cornea (amongst other surfaces). The reflection of this infrared light on the cornea (the corneal reflection) and the position of the pupil are then registered by the eye tracker. A calibration sequence is subsequently run to transform the position of the corneal reflection and the pupil into the gaze position on the screen. Remote video-based eye-trackers are particularly popular for infant research. They can be positioned in front of the infant and do not interfere with the infant unlike, for example, EEG measurements. In addition, they allow infants to move their head in front of the eye tracker. Finally, remote eye trackers are often easy to operate. As a result, eye trackers are currently common in many infant research facilities (Aslin, 2012; Oakes, 2012). The switch from having observers code infants’ gaze direction to having a machine compute it has two major advantages. First, gaze estimation by an eye tracker is objective. Second, the spatial and temporal resolutions with which gaze direction can be determined are markedly higher for eye tracking than for manual coding of videos.

### 3. *Eye-tracking data quality in infancy*

While eye tracking seems an attractive method for investigating all aspects of infant looking behavior, there are several pitfalls to consider. First, while eye-tracking methodology for adults has received considerable attention (see Holmqvist et al., 2011, for an extensive overview), eye-tracking methodology for infant research has not (see e.g. Wass, Forssman, & Leppänen, 2014). This is particularly important, as infants tend not to behave as adults would in front of an eye-tracker. Adults can usually be instructed on how to behave during an eye-tracking experiment. In addition, their head movements can be restrained by using a chinrest, headrest, or bite bar, thereby limiting the interference of movement with the output of the eye-tracker. Infants, on the other hand, cannot readily be instructed to perform a certain task, to remain still, or to focus on a particular stimulus during calibration. Whether and how the quality of eye-tracking data is affected by the behaviors typically seen in infant research is currently unclear (although see Wass et al., 2014, for some first examples). Second, analyzing and interpreting eye-tracking data is a laborious endeavor (Gredbäck, Johnson, & von Hofsten, 2009). It requires not only knowledge of the physiology of the eye, which may differ between adults and infants, but also of signal processing and the technical limitations of the eye tracker (Oakes, 2010). In the present study we investigate which factors affect data quality in infant eye tracking. In doing so we hope to help researchers achieve higher data quality and include data from more participants, and help manufacturers develop better systems for infant eye-tracking research.

#### **3.1. Data quality and its consequences for eye-tracking data analyses**

How does one judge the quality of eye-tracking data? A prerequisite is to know exactly what eye-tracking data quality is (Holmqvist, Nyström, & Mulvey, 2012; Nyström, Andersson, Holmqvist, & van de Weijer, 2013). Data quality refers to a property of the raw eye-tracker data and is often operationalized in several measures, three of which we describe here: spatial accuracy, spatial precision, and data loss (Holmqvist et al., 2011). We will

### 3.1. Data quality and its consequences for eye-tracking data analyses

define these data quality measures one at a time and discuss the possible consequences for event detection, i.e. identifying eye movements (saccades) and periods where the eye is relatively still (fixations) in the eye-tracker signal. Moreover, we will discuss the possible consequences for the calculation of eye-tracking measures when data quality is low. First it should be noted that the accuracy, precision, and amount of data loss that are observed during a measurement stem from a combination of both participant, environmental, and eye-tracker related factors. The accuracy and precision that can be achieved using a particular eye-tracker are often reported by the manufacturer. These values are usually achieved under optimal conditions (i.e. by using artificial eyes or cooperative adult participants) and are difficult to achieve when conducting eye-tracking research with infants (see Hessels, Cornelissen, Kemner, & Hooge, 2015 for a discussion). Regardless of whether decreased accuracy or precision stems from the eye tracker or the participant, they impose restrictions on what conclusions can be drawn from the eye-tracking data.

Accuracy refers to the systematic offset between the gaze position as reported by the eye-tracker and a known target position. Accuracy is calculated by having participants fixate a target and calculating the offset between this target and the gaze position reported by the eye tracker. Figure 3.1B depicts an example trial in which accuracy might be calculated. The distance between the center of the fixation target, and the mean of the gaze coordinates from the left and right eye is an estimate for the accuracy. Stimuli should be of sufficient size and surrounded by a sufficiently large margin to account for the (in)accuracy during a measurement. If accuracy is low, this might result in decreased total dwell times on area of interests (AOIs) if the margins around the AOIs are too small (Holmqvist et al., 2012).

Precision refers to the inverse of the sample-to-sample error during a recording (Holmqvist et al., 2012). Precision can be computed in a number of ways (Holmqvist et al., 2011), but a common measure for precision is the root mean square (RMS) of the Euclidean distances between samples dur-

### 3. *Eye-tracking data quality in infancy*

ing a period where the eye is still. The cluster of gaze coordinates near the center of the fixation target in Figure 3.1B is an example of an imprecise fixation. Low precision has consequences for the separation of fixations and saccades. If a fixed velocity threshold is used to separate fixations from saccades, low precision (i.e. a higher RMS value) might result in a larger number of samples spuriously exceeding this threshold. The number of fixations may then apparently increase (Wass, Smith, & Johnson, 2013). If a velocity algorithm with a threshold that adapts to the noise level is used, this might result in an overall higher threshold. Small saccades might thereby remain below the velocity threshold and the number of fixations apparently decreases (Holmqvist et al., 2012). When the number of detected fixations changes, so does the average fixation duration associated with these fixations, as multiple fixations are either merged into one fixation, or split into multiple fixations (see Shic, Chawarska, & Scassellati, 2008 for possible consequences thereof). Both accuracy and precision may vary independently in eye-tracking data. Figure 3.1A depicts a schematic overview of the possible combinations of accuracy and precision.

Data loss, the final aspect of data quality, refers to the proportion of valid samples that the eye tracker reports (Nyström et al., 2013). An invalid sample occurs when the eye tracker does not report a gaze position. This might be because a participant is looking outside the tracking area (e.g. away from the screen) or because a participant's eyelids are closed due to a blink. Data loss can, however, also occur when the participant is directed towards the screen and the participant's eyes are open. This might be for a number of reasons, for instance, that the eye-tracker is unable to detect the eyes, the pupil, or the corneal reflection. There are several possible consequences of data loss: Detecting fixations or saccades cannot be accomplished for periods where gaze positions cannot be computed, unless interpolation methods are applied to resolve data loss. Furthermore, if a short period of data loss occurs during a fixation, it might appear as if that fixation is actually two separate, shorter, fixations. In addition, data loss might lead to apparently longer latencies, and increased variability, of gaze shifts towards a target (Wass et al., 2014).

### 3.1. Data quality and its consequences for eye-tracking data analyses

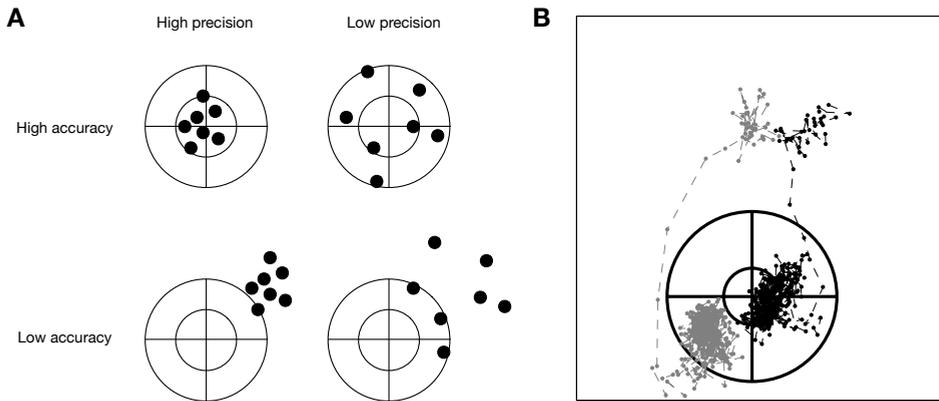


Figure 3.1.: (a) Schematic overview of accuracy and precision, given that the participants gazes at the center of the bulls eye. Dots represent consecutive gaze positions reported by the eye tracker. (b) Example data from a validation trial. Black dots represent data from the right eye; grey dots represent data from the left eye. The bull's eye is a schematic representation of a fixation target.

### 3. *Eye-tracking data quality in infancy*

Although the discussion on eye-tracking data quality has become more prominent over the last few years, reporting data quality measures is still uncommon in the field of infant eye tracking. Current guidelines for publishing infant eye-tracking studies do not specify that measures of accuracy, precision, and data loss achieved during measurement to be included (Oakes, 2010). Inclusion of participants or trials based on data quality is mostly based on the amount of data that was recorded. Chawarska & Shic (2009); Hunnius, de Wit, Vrins, & von Hofsten (2011); Shic, Macari, & Chawarska (2014), for instance, excluded trials based on inattention, and Amso, Haas, & Markant (2014) excluded participants with less than 30% valid data achieved during a recording. In addition, there are studies that specify a minimum accuracy at calibration before proceeding with the study (e.g. S. P. Johnson, Amso, & Slemmer, 2003), or studies implementing post-hoc procedures to correct for inaccurate data (Frank, Vul, & Saxe, 2011). We are unaware of any studies using estimates of precision as exclusion criteria in infant eye-tracking research. The issue of data quality is not only limited to post hoc exclusion or inclusion of data; achieving high quality for the data that are ultimately included is equally important. The question that then arises is how exactly high data quality can be achieved? What are the factors that affect data quality, and what are their relative contributions?

### **3.2. How can higher data quality be achieved?**

Using adult participants, several studies have examined which factors might influence data quality of video-based eye trackers. Kammerer (2009) notes that for participants with ‘bright-colored’ (i.e. bluish color) eyes accuracy was lower compared to participants with darker eyes using a Tobii 1750 remote eye-tracker. In addition, accuracy was lower for participants with glasses compared to participants without glasses or with contact lenses (Kammerer, 2009). Two studies systematically investigated factors that might influence data quality: Nyström et al. (2013) using a SMI HiSpeed 500Hz tower-mounted eye tracker, and Bignaut & Wium (2013) using a re-

### 3.2. How can higher data quality be achieved?

mote Tobii TX300 eye tracker. Nyström et al. (2013) investigated whether calibration method, visual aids, eyelash direction, eye color, the use of mascara, eye dominance, pupil diameter, recording number, and the target position affected accuracy, precision, and data loss for 149 adult participants. They confirm the results by Kammerer (2009) that data quality is lower for participants with bright eye colors compared to participants with dark eye colors. In addition, Nyström et al. (2013) highlight the differences in data quality based on the calibration method used: operator-, participant-, or system-controlled. Blignaut & Wium (2013) report that accuracy and precision are lower for Asian participants compared to African and Caucasian participant. Among others, the operating distance from the eye tracker affected accuracy, precision, and trackability (i.e. the complement to the proportion of data loss), and the vertical head position affected trackability. Finally, Holmqvist et al. (2011) highlight several more specific details of video-based eye trackers that might affect precision and accuracy based on their observations. All in all, there are a large number of factors that may increase or decrease data quality for adult participants.

Eye tracking with infant participants has, however, been less investigated. As Aslin & McMurray (2004) already pointed out, eye tracking with infants can be more difficult than can be expected with adult participants. The differences between eye tracking with infants and eye tracking with adults are most easily observed in raw data. Figure 3.2 depicts an example trial of both an adult and an infant participant. The raw data for the adult participant are precise and the eye tracker continuously reports gaze. The raw data for the infant participant, however, are less precise, and contain periods of data loss due to inattention by the infant and instable tracking of the eyes. The differences between adult and infant eye tracking data may be due to a number of reasons. First, choosing an eye-tracker that is suitable for infant research is a difficult matter, particularly when the technical specifications of an eye tracker do not necessarily predict its ability to cope with infant behavior during measurement (Hessels et al., 2015). Second, seating an infant in front of an eye tracker can be done in numerous ways; we do not yet know how this affects eye-tracking data

### 3. *Eye-tracking data quality in infancy*

quality. Infants can, for instance, be positioned in the lap of the parents (e.g. Gredebäck, Fikke, & Melinder, 2010; Wass & Smith, 2014), in a high chair on the floor, or strapped in a baby seat (e.g. Jones, Carr, & Klin, 2008; Shic et al., 2014). Each type of seating may place different requirements on how the eye tracker can be maneuvered to achieve the right geometry of the setup. In addition, each type of seating may allow more or less movement during a measurement. When possible, infants tend to move around during measurements. This changes both the position and orientation of the eyes relative to the eye tracker (Wass et al., 2014). When movements that are often observed in infant research (i.e. looking away from the screen, and shifted head orientations) were modeled by adult participants, data loss and possibly systematic offsets were observed (Hessels et al., 2015). Notably, the amount of data loss and the severity of possible systematic offsets depended on the eye tracker tested. Third, infants are more difficult to calibrate due to inattention to small calibration targets (Aslin & McMurray, 2004). While larger, moving calibration targets have been used as replacements for infants, little is known about how this might affect the accuracy of measurements. Finally, there might be physiological differences between infants and adults that render eye tracking more difficult. Wass et al. (2014), for example, suggest that an infant's increased pupil size can make tracking problematic, as they are hard to detect for pupil detection algorithms built for adult pupil size. Furthermore, Tobii, a well-known manufacturer of eye trackers in infant research, developed an illumination mode for their TX300 model that is specifically designed for infants. We speculate that this is for differences in the reflection of infrared light between infants' and adults' skin, making detection of the pupil and corneal reflection more troublesome as they become darker in the image when the white balance is adjusted. However, little is known about whether this infant illumination mode increases the data quality or tracking performance of the eye tracker, and how skin reflectance of infrared light affects eye-tracking data quality in general.

We are only aware of one study that has specifically examined data quality in infant eye tracking research. Wass et al. (2014) report a negative

### 3.2. How can higher data quality be achieved?

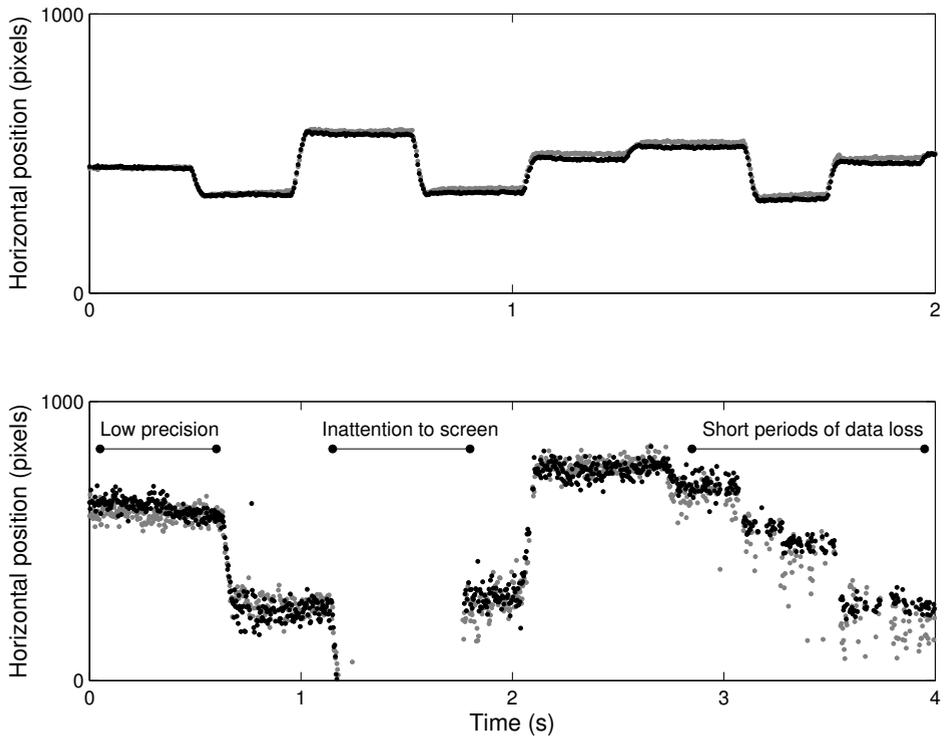


Figure 3.2.: Raw data from an example trial for an adult participant (top panel) and an infant participant (bottom panel) recorded with the Tobii TX300 at 300 Hz. Black dots are coordinates from the right eye; grey dots are coordinates from the left eye. The infant data are less precise than the adult data and contain periods of data loss due to inattention and other acquisition problems. For both adult and infant data, fixations and saccades can still be identified.

### *3. Eye-tracking data quality in infancy*

correlation between amount of head movement and precision, and precision appears to be lower late in a measurement than early in the measurement. In addition, Wass et al. (2014) report that both head movement and time since the start of the measurement are correlated with data loss. They employ the average duration of valid data periods as a measure for data loss: shorter periods of continuous valid data would indicate more data loss. Wass et al. (2014) suggest that this measure better reflects data loss due to unstable tracking of the eyes than data loss due to inattention of the infant. The reasoning behind this is that data loss in infant studies differs from that in adult studies. In adult studies, data loss is often reported as the proportion of invalid samples during a trial or experiment. Usually, the adult participants are instructed to attend to the screen the entire time. If we assume participants follow the instruction, any data loss stems either from blinks or technical difficulties during the measurement. In infant eye-tracking research, however, participants are not expected to attend to the screen the entire time. Data loss can therefore stem not only from blinks or technical difficulties, but also inattention (i.e. looking away from the screen). In order to measure data loss in infant eye-tracking research due to unstable tracking of the eyes, periods of data loss stemming from blinks and inattention should therefore be excluded. We introduce such a measure for data loss in the present study.

Concluding, it would seem that both time since the beginning of a measurement and movement during a measurement decrease data quality in infant eye-tracking research. There have, however, been no endeavors to systematically investigate participant and measurement characteristics that affect eye-tracking data quality with infants. We address this gap in the literature by investigating the effect of predictors we deem most influential on infant eye-tracking data quality in a group of 10-month-old infants who visited our lab twice. Although the purpose of the two visits was to examine test-retest reliability in a visual search task, the present study only concerns data quality assessment. The following predictors were investigated. Eye physiology - the color of the eye, the direction of the eyelashes, and the size of the opening of the eye - is suggested to affect eye-tracking

data quality for adult participants (Blignaut & Wium, 2013; Nyström et al., 2013). For infants, however, the pigment that produces the color of the eye is still being formed; most European infants (i.e. with a light skin color) have bright eyes at birth, with pigment formation occurring throughout the first year after birth. In addition, eyelashes for infants tend to be sparser than for adults. Whether eye-tracking data quality is also affected for infants is a question in the present study. Moreover, where adult participants can be positioned to the researcher's liking and movement can be restricted, the same does not apply for infants. We investigate whether the manner of seating the infant, and the amount of movement it allows, affects data quality. Finally, we wonder whether the behavior and contentedness of the infant during the measurement affects data quality. In infant research (in general, not only eye-tracking research) excluding participants due to 'fussiness' is common, sometimes up to a third of the participants (e.g. Cassia, Turati, & Simion, 2004; Leppänen, Moulson, Vogel-Farley, & Nelson, 2007). Although it is difficult to find specific criteria for excluding fussy infants, common criteria include the infant being upset, hungry, or sleepy. We investigate whether factors related to fussiness also affect eye-tracking data quality. Our findings might help researchers achieve higher data quality and increase the throughput of infants from data recording to data analysis. In addition, our findings might help manufacturers develop better eye-tracking systems for infant eye-tracking research.

## 3.3. Methods

### 3.3.1. Participants

Seventy-seven infants were invited into the lab center for a larger study, recruited through the local municipality. Of the 77 infants invited, 75 (39 male, 36 female) participated in a first session of the present eye-tracking study. Sixty-one (29 male, 32 female) out of the 75 that completed the eye-tracking experiment on their first visit (which we will refer to as a session) returned to the lab center for a second session. A total of 136 sessions were thereby recorded for the present study. Mean age during the

### 3. Eye-tracking data quality in infancy

first session was 302.8 days ( $sd = 12.8$  days); mean age during the second session was 307.5 days ( $sd = 11.2$  days). Infants were only invited to participate if the parents indicated that the infants were not born preterm (i.e. before 37 weeks of pregnancy), had no impaired hearing or vision, or any developmental disorders. Parents gave written informed consent on the day of the first session, and the study was approved by the ethics committee of the local University Medical Centre (Protocol ID 14-221). Parents received a 10 € compensation for each testing day, with another 5 € travel compensation if required.

#### 3.3.2. Operators

As Nyström et al. (2013) report that different operators may achieve varying levels of data quality, we incorporated the operator into our statistical analysis (see statistical analysis for details). Four operators performed the data recordings. Two operators had previous experience with infant eye-tracking recordings using the eye tracker in the present study. One operator had extensive experience with eye tracking in adult participants using multiple systems, but not with infants. The final operator was newly trained and had recorded data in approximately five eye-tracking sessions with infants, and five with adults. The present experiment was part of a larger pilot study in which infants completed a number of tasks across the day. Consequently, the specific time of the day and operator for the eye-tracking sessions were not set in advance; whoever was available performed the data recording. All operators, regardless of experience level, were given the same training in eye tracking with infants by the first author prior to the start of the study.

#### 3.3.3. Apparatus

Stimulus presentation was handled by MATLAB R2013a and the Psych-Toolbox (version 3.0.11; Brainard, 1997) running on a MacBook Pro with OS X 10.9. Stimuli were presented on an external 23-inch screen belonging to the Tobii eye tracker at a resolution of 1920 by 1080 pixels and a refresh rate of 60 Hz. The Tobii TX300 eye tracker running at 300Hz was used

for tracking infants' eye movements. The TX300 is capable of recording at  $0.4^\circ$  accuracy (binocular), and  $0.14^\circ$  precision under ideal conditions. As the accuracy and precision that can be achieved while recording infants is one of the aims of the present study, the actual precision and accuracy values will be presented in the results section. The Tobii SDK was used for communication between MATLAB and the eye-tracker.

### 3.3.4. Stimuli

The experiment consisted of 24 visual search trials (based on Amso & Johnson, 2006), used for the calculation of flicker and RMS noise (see *Data analysis*). Each visual search display consisted of 28 white lines ( $3.3^\circ$  by  $0.9^\circ$ ) as target candidates on a black background. The lines were arranged in a grid of 14 columns by 2 rows, and subsequently jittered between  $-1.6^\circ$  and  $1.6^\circ$  in the horizontal and between  $-6.3^\circ$  and  $6.3^\circ$  in the vertical direction. All lines except the target line were aligned vertically. The target line was tilted  $30^\circ$ ,  $60^\circ$ , or  $90^\circ$  clockwise, and could appear in one of eight fixed locations. Each combination of target line angle and location was presented once, resulting in 24 trials. The visual search trials were interspersed with validation trials after the first and every additional fifth trial for a total of five validation trials. Each validation trial contained one validation target that was identical to one of the calibration targets. The validation targets were used to calculate offset (see data quality measures).

Preceding the visual search experiment was a 5-point calibration sequence. Each calibration and validation stimulus consisted of a colored spiral (red, green, yellow, purple, or blue) on a black background. The spiral changed in size between  $4.0^\circ$  and  $5.4^\circ$  at 0.8 Hz following a sinusoidal wave. In addition, the spiral rotated at 0.8 Hz. Following a key press of the operator, the spiral shrank in size to  $0.5^\circ$  over a period of 0.5 s. The spiral then remained on screen for 0.2 s. For the calibration sequence, a point was calibrated at the start of this 0.2-second period. For the validation trials, data was recorded throughout (see Accuracy and precision for more details).

### *3. Eye-tracking data quality in infancy*

#### **3.3.5. Procedure**

##### **Positioning**

The infants and parents were welcomed into the eye-tracking room and familiarized with the experimental setup. Thereafter, the infants were strapped in a baby seat, and the parent was seated on a height-adjustable chair. The baby seat was subsequently placed on the parent lap, with the infant placed parallel to the screen of the eye tracker. Positioning the infant in a baby seat was done as this would give the most stable positioning through the recording and limit the infants' movements. If, however, the parent indicated that the baby seat would probably result in a restless or upset infant, the infant was seated without a baby seat in the parents lap or in a high chair. The decision for either the parents lap or the high chair was up to the judgment of the operator, i.e., which of the two would work best for the particular infant. While this manner of seating introduces a selection bias for the conditions other than baby seat, the seating predictor is included for the following reason: if data quality is reduced after choosing a different type of positioning, this is important to consider when conducting eye-tracking research in infancy, regardless of whether the seating itself is the (only) cause of this reduced data quality. We outline this issue further in the discussion. After positioning the parent and infant, the position of the eye tracker was adjusted so that the eyes of the infant were at 65 cm from the eye tracker and at the same height as the center of the screen.

##### **Calibration and experiment**

After positioning, a 5-point calibration sequence was started. Calibration stimuli were serially presented in the four corners and center of the screen. The order of points was random each time the calibration was run. The infant was monitored with a webcam. The operator judged from this video whether the infant looked in the direction of the calibration stimulus and pressed the spacebar to calibrate the current point. After the calibration sequence the calibration output was examined. As the Tobii SDK does not provide an objective measure for the accuracy of a calibration, the calibra-

tion output was examined for two features. Calibration points that were either without data, or with data that were inconsistent and characterized by dispersed gaze points around the calibration point were re-calibrated by the operator. Each re-calibration was noted down as an additional calibration run. After calibration was deemed successful, or when the infant started losing attention, the experiment was initiated. A central static attention getter (i.e. a colorful picture) preceded each visual search trial. The operator initiated the trial by pressing the spacebar when the infant was judged to look at the screen. After the first, and each additional fifth trial, a validation target was presented at one of the five calibration locations. When the operator judged from the video that the infant looked in the direction of the validation target, the spacebar was pressed and the validation point shrank in size (see the section Stimuli). If the infant did not attend the validation target for whatever reason, the operator pressed the spacebar as well. If, during the experiment, the infant was not attending to the screen, the operator could present attention getting sounds, or videos (paired with sound) in the center of the screen. The entire experiment, including calibration and positioning, lasted approximately 10 to 15 minutes.

#### 3.3.6. Collected predictors

The predictors that were hypothesized to influence eye-tracking data quality were recorded for all infants where possible. The predictors were divided into three groups: eye physiology, measurement characteristics, and infant contentedness and behavior. The predictors that were included in the statistical analysis are given in Table 3.1 with the number of levels and data points that were available for each predictor.

For eye physiology, the eye color, direction of eyelashes, and size of the eye opening were determined. The operator made the initial assessment of eye color, and pictures were taken to allow a second assessment. Eye color was scored as either ‘bluish’ or ‘non-blue’, to match previous research (Nyström et al., 2013). The eyelash direction was scored as either ‘upward’

### 3. Eye-tracking data quality in infancy

Table 3.1.: Predictors included in the statistical analysis

Predictor	Values
Intercept	Not a predictor. Is given in the results of the statistical model. The intercept of the model represents the value for a theoretical group of infants having the reference value for all categorical predictors, and all numerical predictors set to zero.
Eye color	<b>Bluish</b> (45), non-blue (26), could not be determined for four participants
Seating	<b>Baby seat in parents lap</b> (111), directly in parents lap (15), in high chair (10)
Number of calibrations	Numerical value between 1-4, see collected variables for details
Movement	Numerical value between 0-4, see collected variables for details
Time since fed	Numerical value in hours
Time since awoken	Numerical value in hours
Trial	Numerical value indicating trial number. 1-5 for accuracy. 1-24 for precision and flicker.

**Bold-faced text indicates reference-level of categorical predictors.** Values in parentheses are number of instances in dataset for value of the predictor.

or ‘downward’, and the size of eye opening as ‘open’ or ‘narrow’. After assessment of eyelash direction and size of the eye opening there were five or less instances out of all participants in which ‘downward’ or ‘narrow’ was scored. Both eyelash direction and size of the eye opening were therefore not included in the analysis.

The measurement was characterized by three factors: number of calibrations, seating, and time since the start of the measurement. Number of calibrations was scored between one and four, where a value of four indicated that more than three calibrations sequences were run. Seating was noted as either ‘baby seat in parents lap’, ‘directly in parents lap’, or ‘in high chair’. Finally, a predictor for trial was included to investigate the development of data quality over time since the start of the experiment.

To characterize infant contentedness and behavior during the experiment, three separate measures were collected. The time since the infant was last fed and last awoken was collected. This was done by having the parent report the time the infant last ate and woke up. If infants slept or ate during the time that they were in the lab center, this was noted down. Finally, time since last fed and last awoken was retrieved in minutes by calculating the time difference to the start of the experiment. In addition, the operator scored the overall movement of the infant during the experiment. Operators were instructed to monitor the infant movement throughout the experiment, and estimate the percentage of the total time in the experiment that the infant moved. Movement was scored between zero and four. Zero indicated that the infant did not move noticeably during the experiment. A score of one indicated the infants moved between 0 and 25 percent of the time, a score of two between 25 and 50 percent, three between 50 and 75 percent, and four between 75 and 100 percent. As the present coding scheme was used in prior studies, operators were familiar with this coding scheme as an estimate for movement prior to the start of the present study.

### 3. Eye-tracking data quality in infancy

#### 3.3.7. Data analysis

In order to investigate the influence of the suggested predictors on eye-tracking data quality (i.e. measures for accuracy, precision and data loss), several data reduction steps were taken. First, periods of data loss were extracted from the raw eye-tracking data for the data quality measure of data loss. Second, raw eye-tracking data were reduced into fixations. Third, data quality measures for accuracy and precision were extracted. Finally, statistical models were applied to the three data quality measures. We will outline the separate steps in more detail below.

#### Data loss

Based on experience with infant eye-tracking data, and previous research (Wass et al., 2014), we assumed that data loss due to unstable tracking of the infants' eyes by the eye-tracker (sometimes referred to as technical difficulties) was mainly represented in short periods of data loss. We therefore only investigated data loss where its duration was below 100 ms. We chose this cutoff to exclude periods of data loss due to blinks (roughly 100-400 ms) or periods of inattention to the screen (typically longer than 400 ms). As visible from Figure 3.3, roughly 90% of all periods of data loss observed in the present study had durations of 100 ms or less. We will henceforth refer to these periods of data loss shorter than 100 ms as flicker. The proportion of flicker was calculated by dividing the total duration of flicker in a trial by the total duration of flicker and total duration of all valid samples (i.e. when a gaze position is reported by the eye tracker) combined. We thereby excluded all periods of data loss above 100 ms in the calculation of this measure. For the example of eye-tracking data in Figure 3.2, the measure would be as follows: the sum of all data loss noted under 'short periods of data loss', divided by the amount of data points available (including the short periods of data loss). The longer period of data loss noted under 'inattention to screen' is not included in this measure. For each infant, up to 24 values of proportion flicker were obtained in this manner, one for each of possible 24 visual search trials.

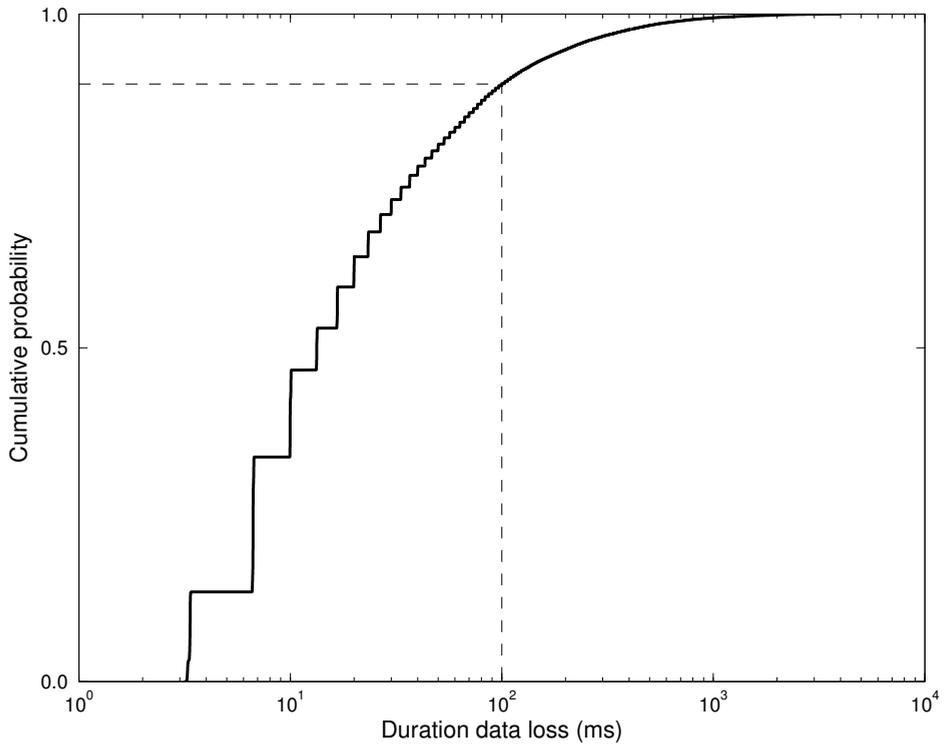


Figure 3.3.: Cumulative probability of the duration of data loss periods in the present study. The dashed line indicates the probability of data loss period occurrences of 100 ms or lower. The steps in the graph correspond to the intersample interval (3.33 ms for 300 Hz data): Data loss duration is a multiple of the intersample interval.

### 3. *Eye-tracking data quality in infancy*

#### **Eye-tracking data reduction**

Raw position signals from the left and right eye were first combined into an average position signal. If gaze position was only available from one eye, that signal was used. Hereafter, a fixation detection algorithm specifically designed for use in infant data was applied. The algorithm operates as an adaptive dispersion algorithm, with which fixation detection can be achieved across larger variations in noise levels, both local and between participants or trials. The algorithm, Identification by 2-Means Clustering (I2MC), is based on a procedure called k-means clustering (where  $k = 2$ ), which is used to determine whether one or two fixation clusters are present in a small moving window. As the I2MC algorithm employs a moving window in which clustering is carried out, it is robust to variations in local noise. In the present study we used a moving window of 200 ms.

#### **Accuracy and precision**

Accuracy was estimated by determining the offset between the center of the validation target and the fixation closest in space to this center. In order to increase the reliability of the fixation position chosen for determining offset, cubic spline interpolation was performed prior to the event detection. Interpolation was done for periods of data loss with a duration of less than 100 ms (Frank, Vul, & Johnson, 2009). This interpolation increased the robustness of fixation detection when short bursts of data loss could occur, thus making the calculation of offset possibly more reliable in the presence of data loss. As there were five validation trials, a maximum of five estimates of accuracy per infant were obtained. Note that offset is a measure for accuracy, and a higher offset means a lower accuracy.

Precision was estimated by calculating RMS noise for all fixations during the 24 visual search trials. Note that no interpolation was done for the estimates of precision, in order to ensure that the RMS noise was not affected or determined by the interpolation method chosen. The number of estimates of precision acquired for each infant depends on the number of fixations made throughout the entire visual search experiment. Note that

RMS noise is a measure for precision and a higher RMS noise means a lower precision.

Histograms for offset, RMS noise, and proportion flicker are presented in Figure 3.4. As the distributions for offset and RMS noise were highly skewed and contained a number of large values, a log-transformation was applied. As visible from Figure 3.4, there are a number of large values that have been obtained for offset. While these values may seem absurdly large, we cannot be sure that these values are incorrect. As described above, the validation stimulus was already  $5.4^\circ$  at its largest, and the infant may have looked anywhere on or around this target. We will therefore not exclude high offset values, and discuss this in more details in the discussion.

### **Statistical analysis**

The influence of the predictors given in Table 3.1 on data quality measures for accuracy, precision, and data loss was statistically tested by fitting linear mixed-effects models using the `lme4` package in R (Bates, Maechler, Bolker, & Walker, 2014; R Core Team, 2014). Linear mixed-effects models (LMEMs) are particularly suited for analyzing data with repeated measurements on both continuous and categorical variables. In addition, LMEMs are superior compared to common statistical analyses (e.g. ANOVAs) when dealing with unbalanced and missing data (Baayen, Davidson, & Bates, 2008).

For all three data quality measures, LMEMs were constructed with operator and participant-session as random effects with random intercepts, and the predictors in Table 3.1 as fixed effects. Random factors are recommended to have at least six levels, and we have only four levels. To test whether this was a problem we tested all predictors using operators as a fixed effect, both as a main effect and an interaction with the other predictors, and we found that the operators had little effect on the predictors of interest. So in order to simplify the models, and because we were not interested in the operators per se, we modeled the operators as

### 3. Eye-tracking data quality in infancy

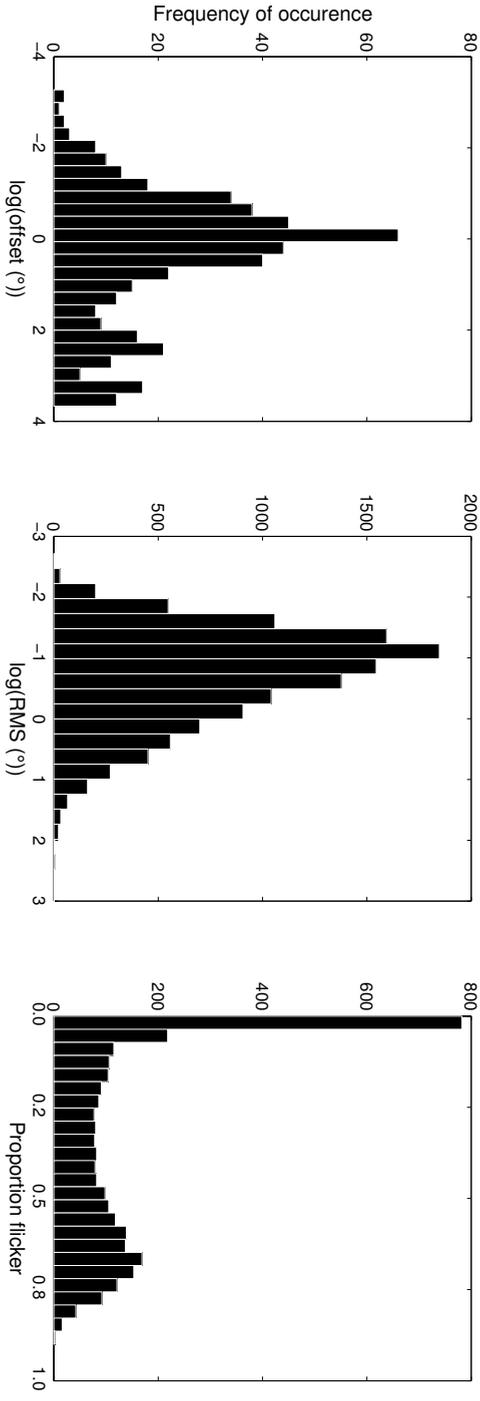


Figure 3.4.: Histograms for the log-transformed offset (estimate for accuracy), log-transformed RMS noise (estimate for precision), and proportion flicker (measure of data loss).

random intercepts. As mentioned, the distributions of offset, RMS-noise, and proportion flicker were skewed. Log transformations were applied to offset and RMS noise to acquire Gaussian-like distributions. In addition a logit transformation was applied to the proportion of flicker. Similar transformations were applied in a previous data quality study with adult participants (Nyström et al., 2013). Estimates for the effect of predictors on the data quality measures were acquired by constructing models with the largest maximal random effects structure that would converge (Barr, Levy, Scheepers, & Tily, 2013): The effect of each predictor was modeled as a fixed effect, and also as a random slope for each participant to account for the variance in effect for each participant. Additionally, the variance of the participants was modeled as random intercepts, and intercept-slope correlations were used to capture any interaction between the participants' intercepts and the effect of the predictor. These models also had operator as random intercepts. If the model failed to converge, the model was simplified to a model with random intercept and uncorrelated slope for that predictor. If that model too failed to converge, the estimate was acquired from the full model with only random intercepts for operator and participant-session. P-values for statistical significance of each predictor were acquired by comparing the full model with all predictors and random intercepts for participant-session and operator, to the model without the predictor of interest using parametric bootstrapping (Halekoh & Højsgaard, 2014). The alpha level was set at 0.05.

As the intercept and estimates reflect log or logit transformed values, the values should be transformed back in order to be interpreted in their respective units (degrees or proportions). The predictors in the model were treatment-coded: An estimate for the accuracy of a new participant given a set of predictors can be acquired by using Equation 3.1, where  $x_n$  is the assigned value in the model (see Table 3.1) for a predictor and  $b_n$  the estimated coefficient for that predictor. The intercept is noted as  $a$ .

$$\log(\textit{accuracy}) = a + b_1 * x_1 + b_2 * x_2 + \dots + b_n * x_n \quad (3.1)$$

### 3. Eye-tracking data quality in infancy

The value can subsequently be transformed back to degrees by taking the exponent.

## 3.4. Results

### 3.4.1. Collected data

Before we turn to the statistical analyses, we first address the amount of data that was observed over the course of the experiment. In infant research, the amount of data collected in the beginning of an experiment is commonly larger than the amount of data collected at the end of the experiment, when more infants have dropped out due to, e.g., inattention or sleepiness. We first examined the extent to which that is the case here. For 124 out of 136 sessions at least one observation for accuracy was acquired. The mean number of observations for accuracy (i.e. fixated validation targets) per session was 3.8 ( $sd = 1.5$ ) out of a maximum five. A total of 472 observations for accuracy were available.

For 134 out of 136 sessions at least one observation for precision was acquired. As visible from the left panel of Figure 3.5, the number of sessions for which at least one observation was acquired decreased as a function of trial number. The linear fit of number of sessions with at least one data point over trials decreased from 102 sessions at trial 1 to 85 sessions at trial 24. The mean total number of observations (i.e. fixations) per session was 101.0 ( $sd = 73.3$ ). A total of 13,530 observations for precision were available. As visible in the middle panel of Figure 3.5, the number of observations also decreased as a function of trial number. The linear fit of number of observations over trials decreased from roughly 650 fixations for trial 1 to roughly 475 fixations for trial 24. This decrease in number of fixations as a function of trial was not due to fixation duration increasing as a function of trial.

For 135 out of 136 sessions at least one observation for flicker was acquired, and, as visible from the left panel of Figure 3.5, the number of

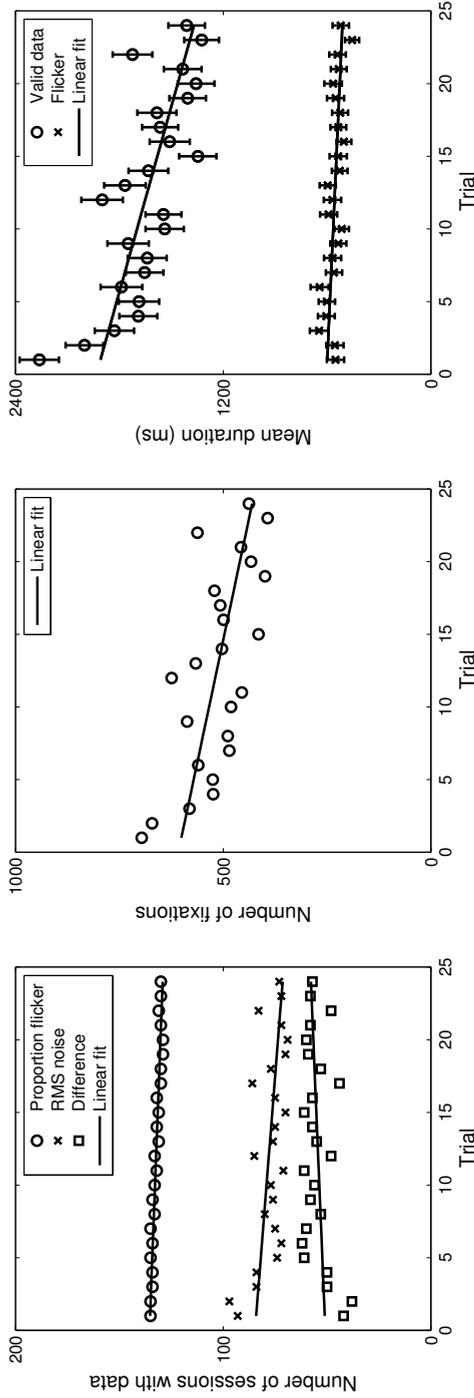


Figure 3.5.: Number of data points and duration of data available as a function of trial. Left panel: number of sessions in which a data point was available for the proportion of flicker, RMS noise, and the difference between these two. Middle panel: number of fixations available per trial; each fixation is one estimate of precision in the linear mixed-effects model. Right panel: mean duration of valid data and mean duration of flicker (i.e., sum of periods of data loss  $< 100$  ms) per trial. Error bars depict standard error of the mean.

### 3. *Eye-tracking data quality in infancy*

sessions for which flicker could be calculated remained fairly stable as a function of trial number. Mean number of observations per session for flicker was 23.3 ( $sd = 2.6$ ) out of a maximum of 24 trials. A total of 3,172 observations for flicker were available.

We will later discuss the implications of the changes in amount of data available over time.

#### **3.4.2. Linear mixed-effects models for offset, RMS-noise, and proportion of flicker**

##### **Offset**

As summarized in Table 3.2, infants with non-blue eyes produced significantly lower offsets than infants with bluish eyes. Moreover, infants seated either directly in the parents lap or in the high chair produced significantly higher offsets than did infants seated in the baby seat on the parents lap. Finally, as the number of calibrations increased, so did the offset for that particular recording session. The total movement during the recording as scored by the operator did not produce higher offsets. Neither did the time since the infant awoke, was fed, or the validation trial number - a measure for time since the start of the experiment.

##### **RMS noise**

As summarized in Table 3.2, infants with non-blue eyes produced significantly lower RMS noise than infants with bluish eyes did. Moreover, the RMS noise increased significantly as a function of trial number. The seating of the infant, the number of calibrations, the total movement during the experiment, and the time since the infant awoke or was fed did not predict RMS noise significantly better than chance.

##### **Proportion of flicker**

As summarized in Table 3.2, infants with non-blue eyes produced a significantly lower proportion of flicker than did infants with bluish eyes. In

Table 3.2.: Statistical results from the linear mixed-effects models for offset, RMS noise, and proportion of flicker.

Predictor	Offset			RMS noise			Proportion of flicker		
	Estimate	SE	p	Estimate	SE	p	Estimate	SE	p
Intercept	0.234	0.384	-	-0.543	0.166	-	-1.028	0.359	-
Non-blue eye color	-0.568	0.184	0.003**	-0.497	0.083	0.001**	-1.429	0.238	0.001**
On parents lap	0.765	0.342	0.022*	0.136	0.153	0.727	0.504	0.438	0.481
In high chair	0.780	0.329	0.022*	-0.026	0.158	0.727	0.354	0.427	0.481
Number of calibrations	0.302	0.089	0.006**	0.034	0.045	0.245	0.231	0.109	0.102
Movement	-0.156	0.091	0.240	0.035	0.042	0.421	-0.074	0.096	1.000
Time since fed	-0.071	0.065	0.514	-0.034	0.042	0.185	-0.107	0.106	0.338
Time since awoken	0.015	0.065	1.000	-0.039	0.030	0.211	-0.044	0.077	0.376
Trial <sup>1</sup>	-0.025	0.041	0.723	0.002	0.002	0.005**	0.008	0.004	< 0.001***

\*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ <sup>1</sup> Note that there was a maximum of five trials for offset, and 24 trials for RMS noise and proportion of flicker

### 3. *Eye-tracking data quality in infancy*

addition, the proportion of flicker increased over trials (as a measure for time since the start of the experiment). No other predictors affected the proportion flicker.

For all models, correlations between fixed effects were unsubstantial to small, except for the correlation between time since fed and time since awoken, which was moderate: Between -0.35 and -0.36 depending on the specific model.

## 3.5. Discussion

The purpose of the present study was to examine factors that might influence data quality in infant eye-tracking research. We investigated eye physiology, measurement characteristics, and infant behavior. Identifying the factors that influence data quality in infant eye-tracking research might help researchers achieve higher data quality and increase the throughput of infants from data recording to data analysis. In addition, our findings might help manufacturers develop better eye-tracking systems for infant eye-tracking research. We showed that eye color, seating, number of calibrations, and trial number, significantly affect data quality measures.

We collected data quality measures for accuracy, precision, and data loss – according to our new definition of data loss, i.e., flicker. First, there were a few observations that could be made from the amount of data available over time. While the number of sessions with at least one observation for precision decreased as a function of trial number, the number of sessions with an observation for flicker did not (see Figure 3.5). This indicates that the number of trials with at least some valid data, which allows calculation of the proportion of flicker, remains stable over the course of the experiment. The number of trials in which event detection can be achieved – which requires more valid data to detect periods of fixation – however, decreases over the course of the experiment. A possible explanation is that inattention to the screen increases as the experiment progresses, leaving little to no data available for event detection in some cases. Another possible explanation is

that infants do look at the screen, but for some reason there is a decrease in stable tracking of the eyes over time. This may cause event detection to become impaired, as there are less consecutive periods of valid data left. Flicker, on the other hand, might still be calculated, as it does not require consecutive periods of valid data, only the presence of valid data. The two explanations are not mutually exclusive.

A first indication that the proportion of flicker per trial increases as a function of trial number comes from Figure 3.5. The right panel of Figure 3.5 shows a decrease of the mean summed duration of valid data (i.e. the sum of valid data periods per trial, averaged over participants) over trials. The mean summed duration of flicker (i.e. the sum of periods of data loss  $< 100$  ms, averaged over participants) remains fairly stable over trial. As the proportion of flicker is defined as the duration of flicker divided by the duration of valid data, we can expect the proportion of flicker to increase over time.

We subsequently modeled whether participant eye physiology, seating, movement, contentedness, and measurement characteristics affected offset (measure for accuracy), RMS noise (measure for precision) and proportion of flicker (measure for data loss due to unstable tracking). We discuss the findings for each eye-tracking data quality measure separately, and then summarize the results for data quality in general.

#### **3.5.1. Accuracy: Offset**

The offset between a known target location and the fixated location by the infants served as our measure for accuracy. When the offset is higher, accuracy is lower, and we henceforth discuss the results in terms of accuracy. Accuracy was significantly higher for infants with non-blue eye color compared to infants with bluish eye color. The finding that bluish color results in lower accuracy has been previously reported by Kammerer (2009) and we extend this finding to infants. In addition, accuracy was significantly lower for participants seated directly in the lap of the parent or high

### 3. *Eye-tracking data quality in infancy*

chair compared to an infant positioned in a baby seat on the parents' lap. We should note, however, that the decision to place an infant directly in the parents' lap or a high chair was not made in advance, and a selection bias was thereby introduced. Not placing an infant in the baby seat on the parents' lap was only done when either the parent indicated that a baby seat would not work, or when the operator determined that the infant would not relax in the baby seat. This means that a lower accuracy for infants in a high chair or directly in the parents' lap might have been a result of external factors. One explanation might be that the willingness of the infant to be restricted in their movement is lower compared to infants who would be positioned in the baby seat. Regardless, the infants who had to be positioned in a different type of seating compared to the baby seat produced lower accuracy. A tentative conclusion might be that a combination of positioning and infant willingness to be restricted in their movement may affect accuracy during recording. Another option may be that the seating outside the baby seat may have allowed infants to shift their position more after calibration. Earlier research has suggested that changing position after calibration affects accuracy in adult eye-tracking research (Cerrolaza, Villanueva, Villanueva, & Cabeza, 2012). Future research with randomized types of positioning will be necessary to substantiate such claims, however.

Finally, accuracy decreased as the number of calibrations before the start of the experiment increased. The intuitive explanation is that the operator gradually reduces the threshold for allowing a calibration. The quality of the final calibration would then be lower than what would be acceptable on a first run. Another explanation might be given if the reasons for re-calibration are considered. When re-calibration was necessary, this was often for one of three reasons: 1) Calibration data for one or more of the points were noisy<sup>2</sup>. 2) Points were not looked at. 3) Data for the bottom or top points were not registered, while the points were actually being looked at. The latter reason usually indicated that the initial positioning of the infant was not adequate. Particularly the angle between eye-tracker and

---

<sup>2</sup>The calibration output is outlined in the manual of the Tobii SDK

the infant had to be adjusted in order to be able to register the missing points. Future research will have to determine whether a gradually decreasing threshold for acceptance of calibration or an incorrect positioning before the calibration sequence produces lower accuracy. One possibility to consider is taking infants who would have to be repositioned out of the baby seat, and do a second positioning after a short break. This might minimize losing the infant's attention when substantial adjustments have to be made to the infant's position.

It is important to note that there are several limitations to the estimation of accuracy here. In adult eye-tracking research, the common method to estimate accuracy is to ask the participant to fixate a target location. The offset between the known target location and the fixated location by the participant then serves as the measure for accuracy. In infant eye-tracking research however, we cannot instruct the participant to fixate a validation target. In order to estimate accuracy it is therefore common to present an attracting stimulus on screen (e.g. a moving picture accompanied by sound), which the infant is assumed to fixate. Subsequently, the stimulus shrinks in size and the offset between the fixated location and the center of the validation target is taken as the estimate for accuracy. Whether the infant follows the shrinking of this validation target has thus far received little attention. In addition, it is difficult to determine which offsets generally occur in an eye-tracking study with infants. In the present study, most of the offsets we observed were below  $2.5^\circ$ , although a number of higher offsets occurred, some even larger than  $10^\circ$ . Note that at its maximum size, the validation target spanned  $5.4^\circ$  (which corresponds to 6.1 cm on the screen). If an infant fixated the edge of the target and remained fixating as the target shrunk, an offset of  $2.7^\circ$  would be recorded. This is when we assume the eye-tracker is recording the infant's gaze with perfect accuracy. As the calibration stimuli were identical to the validation stimuli, the same problem holds for calibration. For the present analysis we used log-transformed offset, which reduced the distance between median values and the extreme values observed in the present study. Most notably, these high offsets were recorded in infants seated directly in the parents' lap or

### 3. *Eye-tracking data quality in infancy*

high chair.

#### **3.5.2. Precision: RMS noise**

The RMS noise in all detected fixations served as estimates for precision. When RMS noise is high, precision is low, and we henceforth discuss the results in term of precision. We included all estimates based on all fixations (i.e. as opposed to the lowest or median value) as infant data may increase or decrease in noise over short periods time. Precision was significantly higher for infants with non-blue eye color compared to infants with bluish eye color. This extends the finding by Nyström et al. (2013), that blue-eyed participants produce noisier data, from adult participants to infants.

Moreover, precision decreased as a function of trial number: precision was lower at the end of the experiment compared to the start. This supports an earlier report by Wass et al. (2014) that precision is higher at the start of the experiment than it is at the end of the experiment. The reason for decreased precision over time might, for instance, be due to changes in position in the head box since the beginning of the experiment or changes in head orientation.

#### **3.5.3. Data loss: proportion of flicker**

The proportion of flicker, calculated by dividing the sum of all periods of data loss shorter than 100 ms by the sum of all valid data in a trial, served as an estimate of data loss due to unstable tracking. A higher proportion of flicker means more data loss, and we henceforth discuss the results in terms of data loss. Data loss was significantly lower for infants with non-blue eye color compared to infants with bluish eye color. While the effect of lighter eye color on estimates of accuracy and precision has been described for adults (Kammerer, 2009; Nyström et al., 2013), and for infants here, no previous research has reported that lighter/bluish eye color produced more data loss. If we consider how decreases in precision and increases in data loss may occur, a possible explanation can be given. As has previously been suggested by Nyström et al. (2013), a possible explanation for blue eyes

producing noisier data is that the blue iris is darker compared to a brown iris under infrared light, making it more difficult to distinguish from the pupil than a brownish iris. If the detection of the pupil is unreliable, this may produce imprecise data, whereas a complete lack of detection produces data loss. If the detection of the eye and pupil is further impaired due to, for example, movement, one might observe more data loss as a result, instead of mere imprecision. Considering that infants are typically more difficult to restrain in terms of movement than adults, we might observe more data loss (in this case specifically flicker) for infants with bluish eye color compared to infants with non-blue eye color. In adults, this might not be the case as detection of the pupil is only slightly impaired due to the eye color, but is not impossible altogether. If more data loss and higher noise is indeed the result of a lower contrast between pupil and iris, than using an eye-tracker that illuminates the eye at a different angle might help. In eye-trackers known as bright-pupil systems (Holmqvist et al., 2011, p. 25), the infrared light is positioned along the same axis as the camera. The infrared light is reflected off the back of the retina, making the pupil appear bright in the camera image. When the pupil is bright and the iris dark under infrared light, it might be easier to determine the pupil-iris border compared to dark-pupil systems. In a bright-pupil eye tracker we therefore expect smaller differences in precision and data loss between eye colors.

Data loss also significantly increased as a function of trial number, which replicates previous reports with infants (Wass et al., 2014). As described before, the mean duration of valid data decreased over time, whereas the mean duration of flicker did not (see Figure 3.5). Data loss (as measured by the proportion of flicker) therefore increased over time.

#### **3.5.4. Summary of eye-tracking data quality models**

We report that for all three measures of data quality, infants with non-blue eye color produced data with higher quality compared to infants with bluish eyes. We hereby extend the finding that eye color influences data

### 3. *Eye-tracking data quality in infancy*

quality in adults (Kammerer, 2009; Nyström et al., 2013) to infants. As in the present study, both previous studies also used a dark pupil eye tracker, where the contrast between a bluish iris and the pupil under infrared light is supposedly low (Nyström et al., 2013), thereby impairing data quality. This is an important finding when, for instance, testing-resources are limited. In this case, one might consider including mainly infants with darker eye colors. However, this is not a guarantee for adequate data quality, and moreover, the researcher should verify that eye color itself is not correlated with the outcome measure of interest. Moreover, both precision and stable tracking of the eyes are impaired as the experiment progresses, consistent with earlier reports (Wass et al., 2014). In addition, we provide possible explanations of how positioning may affect accuracy, although these remain speculative. What about the other predictors we investigated? While eye physiology and measurement characteristics affected data quality, it appeared as though infant contentedness and total movement during the experiment had little effect on data quality. What might be the reasons thereof?

Infant contentedness was operationalized using two measures: the time since the infant was last fed and the time since the infant last awoke. The reasoning was that infants who had not slept or eaten for a longer time were more likely to become fussy during the experiment. We found no effect of infant contentedness on all three data quality measurements. This finding is interesting for two reasons. Given that there was enough variability in the time since infants were last fed or last awoke, one option is that infant contentedness can be considered self-regulatory, where infants who are tired or hungry, cry and are fed or fall asleep. In this manner, infants who start the experiment are neither sufficiently hungry nor sleepy, and unlikely to produce data with impaired data quality. A second possibility is that infant contentedness has, in fact, little to no effect on data quality, which is at least the case for the range that we have collected data in. Regardless of which might be the case, one should not necessarily have to expect impaired data quality due to hunger or sleepiness provided that parents are left to feed their infants, and infants left to sleep at will.

An estimate for overall movement of the infant was obtained by having the operator score the percentage of time the infant moved during the experiment. This estimate of movement did not seem to affect any of the three data quality measures we investigated. Previous research, however, suggested that head movement – calculated from the velocity of eye position in the eye-tracker head box – was negatively correlated with precision, and positively correlated with data loss (Wass et al., 2014). We consider several possibilities that might explain these discrepancies. First, it is quite possible that head velocity per trial predicts lower precision and more data loss (Wass et al., 2014), whereas a crude overall estimate of movement does not (c.f. the present study). An estimate for head velocity from the eye-tracker signal is objective in the sense that it is measured by a machine, whereas an estimate of total movement by the operator is subjective. It might very well be that the estimate by the eye tracker is a better predictor. In the present study, however, operator-coded movement was initially opted for as an estimate for overall movement that would be independent from the signal from the eye-tracker. Independence from the eye tracker was particularly important as we were using it to predict another characteristic of the eye-tracker signal. Head velocity calculated from the eye position signal is not, however, independent from the eye tracker. In essence, using the head velocity signal from the eye-tracker to predict changes calculated from the position signal in the eye-tracker is potentially circular: the eye-tracker signal is used to predict the eye-tracker signal. If this is the case, then the head velocity signal should be affected by the same predictors as the measures for accuracy, precision, and data loss are.

To examine the effect of the predictors that affected data quality in the present study on head velocity, we constructed scatterplots for head velocity calculated as in Wass et al. (2014) against eye color, type of seat, and number of calibrations (see panels A-C Figure 3.6). As visible in panel A in Figure 3.6, infants with bluish eye color seemed to produce larger head velocity compared to infants with non-blue eye color. In addition, as visible from panel B in Figure 3.6, infants seated directly in the parents' lap or in the high chair appeared to have larger head velocities than infants placed

### 3. *Eye-tracking data quality in infancy*

in the baby seat. While the latter might not seem surprising - infants in a baby seat are strapped in and restricted in their movement - the finding that eye color predicts differences in head velocity did surprise us. While there is a possibility that all the infants with bluish eye color generally moved more in our sample than did infants with non-blue eye color, this appeared not to be the case if we compared operator-coded movement for the two groups.

We considered an alternative we deem more plausible. Bluish eye color has been suggested to be more difficult to track robustly due to that fact that the pupil is more difficult to detect in the infrared eye image where the bluish iris appears darker than a brownish iris. The head velocity is calculated from the position signal of the eyes in the head box, and if the same pupil detection method used to determine gaze position is used to determine eye position it should suffer from the same difficulties. Bluish eye colors should therefore result in noisier head position signals calculated from the position of the eyes in the eye-tracker head box. Moreover, the differences between level means we observe in panels A to C in Figure 3.6 correspond to the direction of estimates we observe in the our statistical models: infants with bluish eyes produced lower accuracy (higher offset), lower precision (higher RMS noise), and more data loss. In addition, infants seated in the parents' lap or high chair produced lower accuracy (higher offsets) than infants seated in the baby seat.

Operator-coded movement did show a positive relation with head velocity as calculated from the eye tracker (panel D in Figure 3.6), although it is difficult to determine whether this only reflects general impaired tracking or also actual movement. While overall operator-coded movement did not seem to affect data quality, this estimate is not objective and is crude. It is therefore premature to conclude that overall movement has little effect on eye-tracking data quality. Future research using objective measures of head movement, independent from the eye-tracker signal, will be necessary to fully understand the effects of movement on data quality. It might very well be that overall movement in a 5-minute period does not affect data

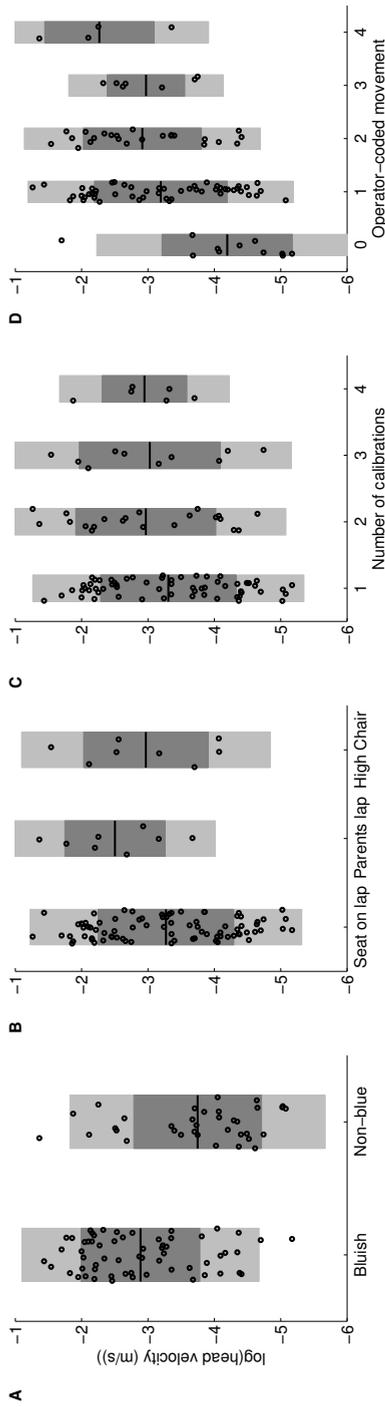


Figure 3.6.: Raw data points for log-transformed head velocity (m/s). Each data point represents the participant mean of head velocities over all 24 visual search trials. Values are jittered with respect to the x-axis to avoid overlapping data points. Thick black lines indicate means, the dark gray area covers one standard deviation from the mean, and the light gray area covers two standard deviations from the mean. (a) Head velocity for infant with bluish and nonblue eye color. (b) Head velocity for different seats used during sessions. (c) Head velocity as a function of number of calibrations. (d) Head velocity as a function of operator-coded movement.

### 3. *Eye-tracking data quality in infancy*

quality, whereas short bursts of movement, which may not be reflected in our overall measure of movement, temporarily decrease data quality (see also Hessels et al., 2015). Head movement in future research might best be estimated using 3d-accelerometers, EMG, or video coding. If the latter is opted for, a separate video system separate from the eye tracker needs to be implemented (particularly given that the Tobii TX300 does not provide a video signal), and particular care should be taken to ensure adequate video quality to code head movement. At present, caution is advised when head movement estimated from the eye-tracker signal is used to predict other aspects of the eye-tracker signal.

## **3.6. Conclusions, limitations, and future research**

In the present study we have investigated the influence of eye physiology, positioning, measurement characteristics, and infant contentedness and behavior on eye-tracking data quality in infant research. We report that eye color influenced all measures of data quality, and precision decreased and data loss increased as a function of trial number. Furthermore, we presented tentative arguments on the effect of positioning on accuracy, although future research is needed to substantiate the present findings. Moreover, we report that infant contentedness during measurement does not seem to affect data quality. Finally, we reported that operator-coded total movement during the experiment did not affect data quality, and highlighted the problem when using head movement estimates from the eye-tracker signal. These findings are valuable for infant eye-tracking researchers, both for the interpretation of data as well as for data collection. Researchers may use the present findings to increase throughput in their eye-tracking studies and optimize data collection. In addition, researchers should be aware of the differences in data quality that can arise due to factors such as eye color. If, for instance, two groups are being compared on measures that are affected by the level of data quality (see e.g. Shic et al., 2008), it is important to balance factors that are known to affect data quality over these groups, e.g. eye color. Finally, manufacturers might use the present results to improve

### 3.6. Conclusions, limitations, and future research

eye-tracking systems for use with infant participants.

There are also several limitations of the present study to consider. First, the eye-movement data presented here were recorded using a Tobii TX300. While this eye-tracker is a common eye tracker for use in infant research, the question remains how the present findings generalize to other systems. In addition, we wonder how the present study extrapolates over different age groups. While some findings we report here are similar as in adults, and are likely to generalize over children, other findings are, at this point, still specific to the age group reported here. Finally, we should note that the Tobii TX300, as well as the eye-trackers used in previous data quality studies (Kammerer, 2009; Nyström et al., 2013; Wass et al., 2014) are dark-pupil eye trackers. When a different eye illumination angle is used, such as that in bright-pupil eye trackers, we expect data quality to be differently affected by some of the predictors we investigated. Particularly, we expect the effect of eye color on data quality measures to be reduced, as the plausible explanation of low contrast between iris and pupil does not hold for bright-pupil systems (Nyström et al., 2013). We encourage and welcome researchers to investigate eye-tracking data quality for other systems, as well as different age groups. Other limitations in the present study that have already been discussed above were the absence of random assignment of the positioning during the experiment and the lack of estimates of movement at a higher temporal resolution (i.e. using motion capture or 3-d accelerometer techniques). We welcome future research into eye-tracking data quality using these techniques.

## Acknowledgements

The authors would like to thank all employees at the KinderKennisCentrum of Utrecht University for help with data collection. In addition, author RH would like to thank Kenneth Holmqvist and the Lund University Humanities lab for their hospitality while working on this project. The study was financed through the Consortium on Individual Development (CID). CID is funded through the Gravitation program of the Dutch Ministry of

### *3. Eye-tracking data quality in infancy*

Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant number 024.001.003 awarded to author CK). Author RA furthermore acknowledges support from the Swedish Research Council, grant no. 437-2014-6735.

## References

- Amso, D., & Johnson, S. P. (2006). Learning by selection: Visual search and object perception in young infants. *Developmental Psychology*, 42(6):1236–1245.
- Amso, D., Haas, S., & Markant, J. (2014). An eye tracking investigation of developmental change in bottom-up attention orienting to faces in cluttered natural scenes. *PLOS One*, 9(1):e85701.
- Aslin, R. N. (2007). What's in a look? *Developmental Science*, 10(1):48–53.
- Aslin, R. N. (2012). Infant eyes: A window on cognitive development. *Infancy*, 17(1):126–140.
- Aslin, R. N., & McMurray, B. (2004). Automated corneal-reflection eye tracking in infancy: Methodological developments and applications to cognition. *Infancy*, 6(2):155–163.
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, 59(4):390–412.
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, 68(3):255–278.
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. R package version 1. Retrieved from <http://CRAN.R-project.org/package=lme4>.
- Bligaaut, P., & Wium, D. (2013). Eye-tracking data quality as affected by ethnicity and experimental design. *Behavior Research Methods*, 46(1):67–80.
- Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, 10(4):433–436.
- Cassia, V. M., Turati, C., & Simion, F. (2004). Can a nonspecific bias toward top-heavy patterns explain newborns' face preference? *Psychological Science*, 15:379–383.
- Cerrolaza, J. J., Villanueva, A., Villanueva, M., & Cabeza, R. (2012). Error characterization and compensation in eye tracking systems. *Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '12*, 205–208.
- Chawarska, K., & Shic, F. (2009). Looking but not seeing: Atypical visual scanning and recognition of faces in 2 and 4-year-old children with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, 39(12):1663–1672.
- Frank, M. C., Vul, E., & Johnson, S. P. (2009). Development of infants' attention to faces during the first year. *Cognition*, 110(2):160–170.

### 3. Eye-tracking data quality in infancy

- Frank, M. C., Vul, E., & Saxe, R. (2011). Measuring the development of social attention using free-viewing. *Infancy*, 17(4):355–375.
- Gredebäck, G., Fikke, L., & Melinder, A. (2010). The development of joint visual attention: a longitudinal study of gaze following during interactions with mothers and strangers. *Developmental Science*, 13(6):839–848.
- Gredebäck, G., Johnson, S., & von Hofsten, C. (2009). Eye tracking in infancy research. *Developmental Neuropsychology*, 35(1):1–19.
- Halekoh, U., & Højsgaard, S. (2014). A Kenward-Roger approximation and parametric bootstrap methods for tests in linear mixed models – The R package pbrktest. *Journal of Statistical Software*, 59(9):1–32.
- Hessels, R. S., Cornelissen, T. H. W., Kemner, C., & Hooge, I. T. C. (2015). Qualitative tests of remote eyetracker recovery and performance during head rotation. *Behavior Research Methods*, 47(3):848–859.
- Holmqvist, K., Nyström, M., & Mulvey, F. (2012). Eye tracker data quality: What it is and how to measure it. *Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '12*, 45.
- Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press.
- Hunnius, S., de Wit, T. C. J., Vrans, S., & von Hofsten, C. (2011). Facing threat: Infants' and adults' visual scanning of faces with neutral, happy, sad, angry, and fearful emotional expressions. *Cognition & Emotion*, 25(2):193–205.
- Johnson, S. P., Amso, D., & Slemmer, J. A. (2003). Development of object concepts in infancy: Evidence for early learning in an eye-tracking paradigm. *Proceedings of the National Academy of Sciences*, 100(18):10568–10573.
- Jones, W., & Klin, A. (2013). Attention to eyes is present but in decline in 2–6-month-old infants later diagnosed with autism. *Nature*, 504:427–431.
- Jones, W., Carr, K., & Klin, A. (2008). Absence of preferential looking to the eyes of approaching adults predicts level of social disability in 2-year-old toddlers with autism spectrum disorder. *Archives of General Psychiatry*, 65(8):946–954.
- Kammerer, Y. (2009). How to overcome the inaccuracy of fixation data? The development and evaluation of an offset correction algorithm. *Presented at the Scandinavian Workshop on Applied Eye Tracking 2009*.
- Leppänen, J. M., Moulson, M. C., Vogel-Farley, V. K., & Nelson, C. A. (2007). An ERP study of emotional face processing in the adult and infant brain. *Child Development*, 78(1):232–245.
- Nyström, M., Andersson, R., Holmqvist, K., & van de Weijer, J. (2013). The in-

### 3.6. Conclusions, limitations, and future research

- fluence of calibration method and eye physiology on eyetracking data quality. *Behavior Research Methods*, 45(1):272–288.
- Oakes, L. M. (2010). Infancy guidelines for publishing eye-tracking data. *Infancy*, 15(1):1–5.
- Oakes, L. M. (2012). Advances in eye tracking in infancy research. *Infancy*, 17(1):1–8.
- R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org/>.
- Shic, F., Chawarska, K., & Scassellati, B. (2008). The amorphous fixation measure revisited: With applications to autism. *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*.
- Shic, F., Macari, S., & Chawarska, K. (2014). Speech disturbs face scanning in 6-month-old infants who develop autism spectrum disorder. *Biological Psychiatry*, 75(3):231–237.
- Wass, S. V., & Smith, T. J. (2014). Individual differences in infant oculomotor behavior during the viewing of complex naturalistic scenes. *Infancy*, 19(4):352–384.
- Wass, S. V., Forssman, L., & Leppänen, J. (2014). Robustness and precision: How data quality may influence key dependent variables in infant eye-tracker analyses. *Infancy*, 19(5):427–460.
- Wass, S. V., Smith, T. J., & Johnson, M. H. (2013). Parsing eye-tracking data of variable quality to provide accurate fixation duration estimates in infants and adults. *Behavior Research Methods*, 45(1):229–250.
- Wheeler, A., Anzures, G., Quinn, P. C., Pascalis, O., Omlin, D. S., & Lee, K. (2011). Caucasian infants scan own- and other-race faces differently. *PLOS One*, 6(4):e18621.



## **4. Noise-robust fixation detection in eye-movement data – Identification by two-means clustering (I2MC)**

Published as:

Hessels, R. S., Niehorster, D. C., Kemner, C., & Hooge, I. T. C. (2016). Noise-robust fixation detection in eye-movement data – Identification by two-means clustering (I2MC). *Behavior Research Methods*.

Author contributions:

RH designed the algorithm. DN optimized the algorithm. RH, DN, CK, IH designed the study. RH, DN analyzed the data. RH, DN, IH interpreted the data. RH, DN drafted the paper. RH, DN, CK, IH finalized the paper. RH and DN contributed equally to this work.

## Abstract

Eye-tracking research in infants and older children has gained a lot of momentum over the last decades. Although eye-tracking research in these participant groups has become easier with the advance of the remote eye tracker, this often comes at the cost of poorer data quality than in research with well-trained adults (Hessels, Andersson, Hooge, Nyström, & Kemner *Infancy*, 20, 601–633, 2015; Wass, Forssman, & Leppänen *Infancy*, 19, 427–460, 2014). Current fixation-detection algorithms are not built for data from infants and young children. As a result, some researchers have even turned to hand correction of fixation detection (Saez de Urabain, Johnson, & Smith *Behavior Research Methods*, 47, 53–72, 2015). Here we introduce a fixation-detection algorithm – identification by two-means clustering (I2MC) – built specifically for data across a wide range of noise levels and when periods of data loss may occur. We evaluated the I2MC algorithm against seven state-of-the-art event detection algorithms, and report that the I2MC algorithm’s output is the most robust to high noise and data loss levels. The algorithm is automatic, works offline, and is suitable for eye-tracking data recorded with remote or tower-mounted eye trackers using static stimuli. In addition to application of the I2MC algorithm in eye-tracking research with infants, school children, and certain patient groups, the I2MC algorithm also may be useful when the noise and data loss levels are markedly different between trials, participants, or time points (e.g., longitudinal research).

The emergence of the remote video-based eye-tracker has allowed researchers to conduct eye-movement research with a plethora of participant groups for which conventional eye-tracking techniques are unsuitable. Unlike, for example, scleral coil techniques or head-mounted and tower-mounted video-based eye trackers, remote eye-trackers can be positioned at a distance from the participants and allow them to move freely within a specified range. Remote video-based eye trackers are therefore suitable to use in participant groups where head movement is difficult to restrain, such as infants (Oakes, 2012) or school children (e.g. Holmberg, Holmqvist, & Sandberg, 2015). As a result, eye-tracking research in, for instance, infants has gained a lot of momentum over the last decades (e.g. Aslin & McMurray, 2004; Oakes, 2012). While eye-tracking research in these participants groups has become easier with the advance of the remote eye tracker, and although studies are available that provide advice on how to choose an eye-tracker for research in non-optimal conditions (Hessels, Cornelissen, Kemner, & Hooge, 2015b), data quality is still often low compared to recordings of well-trained adults (Hessels, Andersson, Hooge, Nyström, & Kemner, 2015a; Wass, Forssman, & Leppänen, 2014; Wass, Smith, & Johnson, 2013). Current solutions for automatic detection of one of the most commonly investigated events, fixations, in eye-tracking data are not built for low quality data. This applies to both the solutions provided by eye-tracker manufacturers and the research community. This is problematic for eye-tracking research with infants and young children, where data of low quality frequently occurs. As a result, some researchers have moved away from fully automatic analysis techniques and turned to manual correction of fixation detection in eye-movement data from infants (Saez de Urabain, Johnson, & Smith, 2015). Here we consider the consequences of low data quality for fixation detection, describe and quantify the noise in infant data from remote video-based eye trackers, and introduce a new and superior solution for detecting fixations in noisy data.

In humans, visual acuity is greatest in the fovea, and eye-movements are made to bring an area of the visual scene onto the fovea, or maintain it there. There is a distinction between the periods in which an area

#### 4. Noise-robust fixation detection

of the visual scene is kept on the fovea – a fixation – and periods when an area of the visual scene is brought onto the fovea – a rapid eye-position change called a saccade. Figure 1 (top panel) depicts typical eye-movement data from an eye-tracker using static stimuli. At first glance, the periods where gaze position is constant – the fixations – can clearly be discerned from the periods of rapid gaze-position change – the saccades. Labeling segments of the eye-movement data as fixations and saccades may give researchers insight into the spatiotemporal processing of a visual scene. Algorithms that label eye-movement data in this fashion are referred to as event detection algorithms, where an event can be a fixation, smooth pursuit (when using moving stimuli; e.g. Larsson, Nyström, Andersson, & Stridh, 2015), saccade, blink, post-saccadic oscillation (see e.g. Nyström, Hooge, & Holmqvist, 2013), etc. In the present paper we focus on the labeling of fixations in data from remote or tower-mounted eye trackers using static stimuli. More specifically, we investigate fixation labeling under varying levels of noise in the eye-movement data, to mimic fixation detection in low- and high-quality eye-movement data.

### 4.1. Event detection and data quality

An event-detection algorithm generally consists of two parts. The first part, which we refer to as the “search rule”, aims to separate fast periods (saccades) and slow periods (fixations) in the data from each other. The second part, the “categorization rule(s)”, accepts, rejects, and/or merges the saccade and/or fixation-candidates from the first rule according to a set of criteria. These criteria may be, for instance, a minimum fixation time, maximum saccade duration etc. Moreover, these criteria may be based on physiological constraints of the eye, for example a maximum acceleration during a saccade (Nyström & Holmqvist, 2010), or on the experimental setup, for example a minimum saccade amplitude of  $2^\circ$  when elements to be fixated are spaced  $4^\circ$  apart. Event-detection algorithms are often referred to by their search rule. For example, two popular types of event detection algorithms for labeling fixations and saccades are velocity- and

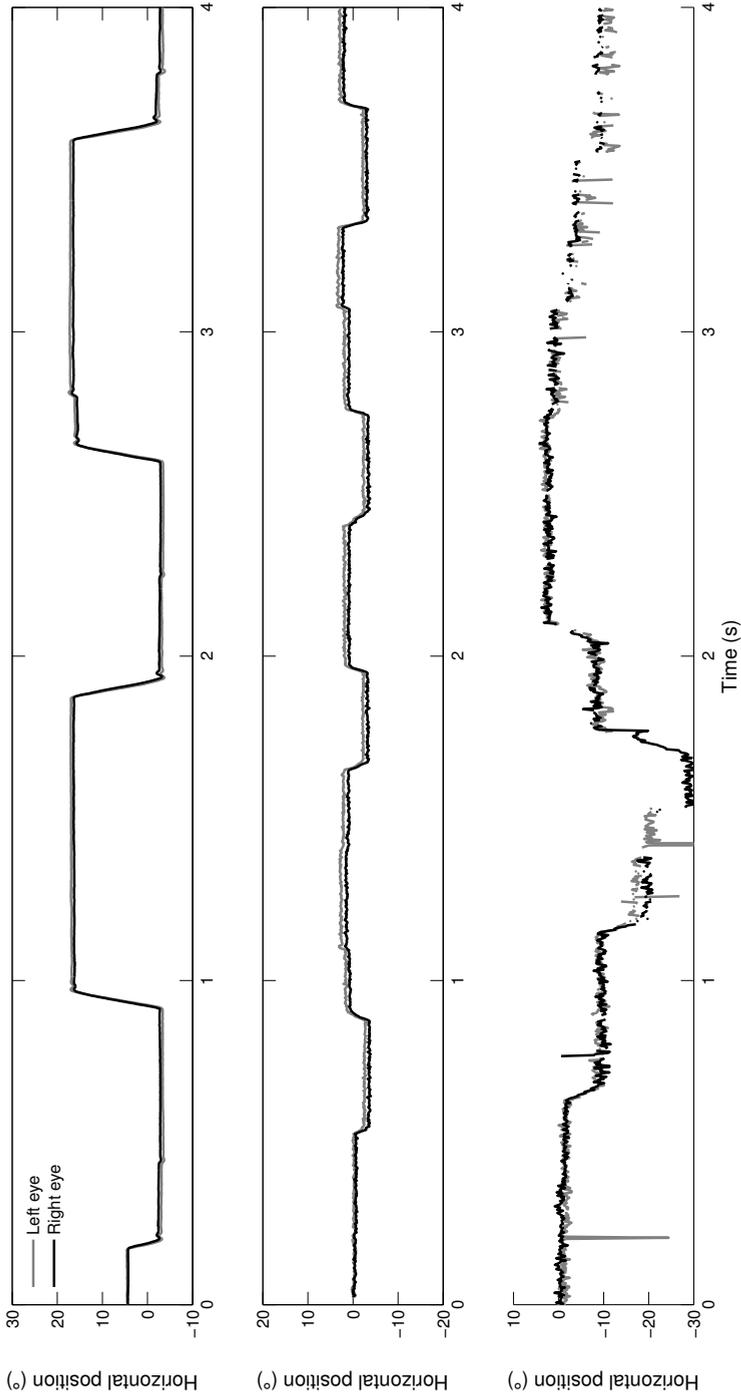


Figure 4.1.: Example eye-tracking data. Top graph depicts data recorded at 500Hz on the SR Research EyeLink 1000 from an adult participant by Hooge et al. (2015). Middle graph depicts data recorded at 300Hz on the Tobii TX300 from an adult participant by Hessels et al. (2016b). Bottom graph depicts data recorded at 300Hz on the Tobii TX300 from an infant participant by Hessels et al. (2015a). Only horizontal coordinates are shown; middle of the screen is at  $0^\circ$ .

#### 4. *Noise-robust fixation detection*

dispersion-based algorithms (see Holmqvist et al., 2011, p. 147–175 for an elaborate overview). Velocity-based event detection algorithms compute a velocity signal from the gaze-position signal, and subsequently use a velocity cut-off to label periods of data as fixation-candidates. The velocity cut-off used may be set in advance, for example at  $30^\circ/\text{s}$ . Other strategies involve first detecting saccade-candidates with a fixed threshold, and subsequently finding fixation start and end points by adapting the threshold to the mean velocity in a period preceding the saccade-candidate (Smeets & Hooge, 2003), or using the median velocity plus a certain number of standard deviations (Engbert & Kliegl, 2003). Dispersion-based algorithms, on the other hand, label periods of data as fixation-candidates when a set of subsequent samples exceed a minimum time and do not exceed a maximum distance from each other. As long as subsequent samples stay within the distance limit, they are added to the fixation-candidate. If a subsequent sample does, however, exceed the maximum distance, the current fixation-candidate is ended and a new fixation-candidate is started in the next available window that fits the minimum time and maximum distance requirements. Both event-detection algorithms may subsequently employ different or identical categorization rules to accept, reject, or merge fixation-candidates into fixations.

The two classes of event-detection algorithms just discussed are widely applied for analyzing eye-movement data. While such event detection algorithms may lead to reasonable results for adult data with large saccades and low noise level (Figure 1, top panel), they may not necessarily do so for data with a larger noise level. Figure 1 (middle panel) depicts adult data in which the noise level is higher and smaller saccades were made. However, fixations and saccades are still relatively easy to distinguish at first glance. In infant data (Figure 1, bottom panel), the amplitude of noise is frequently higher than for adult data, and there are often short bursts in which no data is reported by the eye tracker (Hessels et al., 2015a; Wass et al., 2014). How these differences in data quality affect event detection may depend on the specific event-detection algorithm used.

#### 4.1. Event detection and data quality

Two aspects of data quality are important to consider for the present event detection purposes: spatial precision and data loss<sup>1</sup>. First, the spatial precision of the data refers to the reliability of the measurement when no movement of the eye takes place: the variable error, or noise, in the signal. Although the eye always moves slightly (e.g. tremor or drift), one way to estimate the noise amplitude is to calculate the sample-to-sample position change during fixation (Holmqvist, Nyström, & Mulvey, 2012). When there is a low sample-to-sample position change, the noise amplitude is low. One may observe that the noise amplitude is lowest for the adult data using the SR Research EyeLink 1000 (Figure 1, top panel), followed by the adult data using the Tobii TX300 (Figure 1, middle panel), and finally the infant data using the Tobii TX300 (Figure 1, bottom panel). The noise level is not only determined by the hardware (i.e. the eye tracker), but also by the behavior of the participant group (Holmqvist et al., 2011). For instance, the poorer data quality in infant eye-tracking research may be in part due to the higher amount of movement in infants. An increase in noise amplitude may affect outcome measures such as number of fixation-candidates and mean duration of fixation-candidates. Moreover, depending on the specific search-rule used, the number of fixation-candidates and fixation duration may either increase or decrease. If a fixed velocity threshold is used to separate fixation from saccades, decreased precision may break up long fixation-candidates into shorter fixation-candidates due to noise spuriously exceeding the threshold (Wass et al., 2013). When long fixation candidates are broken up into multiple shorter fixation-candidates, the number of fixations increases and the mean fixation duration decreases. On the other hand, if a velocity threshold is adaptively chosen based on the noise amplitude in the data, small saccades with a velocity close to that of the noise may be missed. This may then cause multiple fixation-candidates to be merged into longer fixation-candidates (Holmqvist et al., 2012). As a consequence, the number of fixations decreases, and the mean fixation duration increases.

---

<sup>1</sup>While accuracy is also an aspect of data quality, this does not affect the detection of fixations.

#### 4. Noise-robust fixation detection

The second aspect of data quality, data loss, refers to periods in which no position coordinates are reported by the eye tracker. While this intuitively may be attributed to a participant not looking at the screen, data loss may often occur due to unstable tracking of the eye by the eye tracker (Hessels et al., 2015a; Wass et al., 2014). Figure 1 (bottom panel) depicts such brief loss of contact: between 3 and 4 seconds there are short periods of recorded data interrupted by short periods of data loss. During event detection, fixation-candidates may be broken up by periods of data loss (Holmqvist et al., 2012). When more data loss occurs, the number of fixations increases, and the mean fixation duration decreases, compared to lower data loss levels. Whether changing the parameters of the categorization rule(s) may compensate the differences in output by the search rule is part of ongoing research (Zemblys & Holmqvist, 2016). On the other hand, while the problem of reduced data quality in, for instance, infant research is a common one, few solutions have been designed to accommodate it.

### 4.2. Event detection in noisy data

To our knowledge, two solutions have been proposed specifically to accomplish event detection in noisy data (Saez de Urabain et al., 2015; Wass et al., 2013). Wass et al. (2013) adapted a velocity-based event detection algorithm specifically designed to cope with eye-movement data from infants. The search rule of this algorithm is as follows: the algorithm first selects only the portions for which data for both eyes were available and applies a smoothing procedure. Subsequently, periods of data loss up to 150 ms are interpolated if the velocity between start and end of the data loss period does not exceed the velocity threshold. Hereafter, all periods of data below the velocity threshold of  $35^\circ/\text{s}$  were marked as fixation-candidates. The categorization rules that follow to label fixation-candidates as fixations are extensive. First, if a fixation-candidate bordered on a period of data loss it was excluded. Second, saccades were excluded (and fixation-candidates consequently merged) if the fixation-candidates before and after were within  $0.25^\circ$  distance from each other. Third, if a saccade is preceded

by a fixation-candidate with an average velocity over  $12^\circ/\text{s}$ , the saccade and bordering fixation-candidate were excluded. Fourth, if a saccade is preceded by three samples with an average velocity over  $12^\circ/\text{s}$ , the saccade and bordering fixation-candidate were also excluded. Fifth, if distance between the gaze positions for the two eyes prior to the saccade was larger than  $3.6^\circ$ , the saccade and bordering fixation-candidates were excluded. Finally, fixation-candidates shorter than 100ms were excluded. Wass et al. (2013) report that this algorithm remains reliable for data containing higher noise amplitude, whereas standard dispersion-based algorithms decrease in reliability with increasing noise amplitude. Saez de Urabain et al. (2015), on the other hand, use a two-step approach to event detection. In a graphical user interface, the user can set a number of parameters for a first estimation of which data segments are fixation-candidates (i.e. the search rule). Hereafter, the user may manually correct the fixation-candidates (i.e. a manual categorization procedure). While the machine-coding, manual-correction approach by Saez de Urabain et al. (2015) may increase the amount of eye-movement data that can be successfully labeled as fixations and used in further analysis, it is a highly time-consuming, and subjective, process. Moreover, whereas the velocity-threshold adaptation by Wass et al. (2013) is automatic, it features a large number of categorization rules for rejecting data, leading to significant amounts of data being excluded when noise level is high. Intuitively, an ideal approach would be an algorithm that is automatic, and reliably achieves fixation labeling in periods of noisy data instead of excluding such data. Here we introduce such an approach.

The present paper introduces the Identification by 2-Means Clustering (I2MC) algorithm. This algorithm was specifically designed to accomplish labeling of fixations across a wide range of noise levels, and when periods of data loss may be present, without the need to set a large number of parameters or perform manual coding. Before we introduce how the algorithm operates, we first discuss how to the algorithm is to be evaluated.

### 4.3. Evaluating the algorithm

How can the output of an event-detection algorithm be evaluated? Or, the question that more generally arises; is the output from my event-detection algorithm “correct”? Intuitively, it makes sense to ask this question. One would want to know whether the fixations labeled by an algorithm are “correct”, or whether the output of the algorithm conforms to a golden standard. The problem is, however, that while researchers appear to informally discuss fixations and saccades with relative ease, there is no golden standard for when a fixation starts or stops (Andersson, Larsson, Holmqvist, Stridh, & Nyström, 2016). Essentially, a fixation is defined by how it is computed, which differs for each event-detection algorithm. Holmqvist et al. (2011) also note *“In reality, perfect matches between the fixations detected by an algorithm and moments of stillness of the eye are very rare. To make matters worse, the term fixation is sometimes also used for the period during which the fixated entity is cognitively processed by the participant. The oculomotor, the algorithmically detected, and the cognitive ‘fixations’ largely overlap, but are not the same”* (p. 150). Tackling the evaluation of algorithms based on the fixations detected is therefore problematic, as the definitions for fixations differ between algorithms.

Komogortsev, Gobert, Jayarathna, Koh, & Gowda (2010), for instance, aimed to determine a goodness-of-fit of the eye-movement data to the stimulus that was presented. Participants were presented with a white dot that appeared sequentially at 15 locations on screen for one second each. In this case, an event-detection algorithm was considered ideal if it detects 15 fixations, 14 saccades, an average fixation duration of one second, and an accuracy of 0°. Importantly, Komogortsev et al. (2010) note that this method is inherently flawed. For instance, eye trackers typically report accuracies of 0.5°, even under ideal circumstances. Moreover, 14 saccades and 15 fixations would imply that no corrective saccades are made. However, Komogortsev et al. (2010) prefer their method over manual techniques, which *“are frequently used to classify eye movement behavior. However, this type of classification technique [i.e. manual coding] is susceptible to*

*human error and can be open for biased interpretation with limited generalizability*” (p. 2643). Andersson et al. (2016), on the other hand, employed two experienced human coders as their golden standard for comparing algorithms to. They note that while human coders generally agreed more with each other than with automatic algorithms, there is a problem. Human coders make mistakes and have disagreements, and eventually it is impossible to determine whether an algorithm or human coder should be “right”. Lastly, in the approach by Zemblys & Holmqvist (2016), the algorithm considered “best” by the authors is taken as the golden standard – a choice that will most likely provoke debate. To sum, there is little consensus on the manner of evaluating the performance of a fixation-detection algorithm.

Instead of focusing on whether a golden standard for the evaluation of fixation-detection algorithms can be approximated, we take a different approach here. The purpose of the presented algorithm, called I2MC, is to achieve consistent labeling of fixations when there may be large differences in data quality between participant, and between trials, as is often encountered in, for instance, infant research. This means that the I2MC algorithm should achieve fixation labeling across a range of noise amplitudes, and when short periods of data loss may be present. If labeling of fixations is done for data with small noise amplitude and few periods of data loss, the output hereof can be compared to the output after adding noise of higher amplitude and periods of data loss to the same data. If the number of labeled fixations, and the corresponding distribution of fixation duration, remain unchanged as noise and data loss increase, the algorithm is considered to be robust to noise and data loss. The key decision left to the eye-movement researcher is then whether the fixation-labeling output at low noise and data loss levels is satisfactory. If one is satisfied with the initial output of the algorithm, one can generalize this satisfaction to higher noise and data loss levels given the algorithm’s stability in the face of increasing noise and data loss. This decision inevitably has to be made by every researcher, as no two experimental setups and data sets are identical.

#### 4. *Noise-robust fixation detection*

In order to evaluate whether the I2MC algorithm improves over current available solutions, 7 competing state-of-the-art algorithms were chosen from the literature. The general motivation for including an algorithm is that it should be able to deal with one or more of the data quality issues that we've outlined above. The specific motivations are given in the Methods section. As discussed briefly, algorithms may differ in both their search rule and categorization rule for labeling fixations. Moreover, the preprocessing steps, such as data smoothing and interpolation to impute missing data, prior to application of these rules may differ between algorithms. It is paramount to note that the focus here is not on finding a combination of pre-processing steps, search and categorization rules that produces the most noise-robust output by testing all possible combinations and exhaustively search through their parameter spaces. Instead, the focus is on whether previous solutions, taken as is, produce noise-robust output, and whether the present I2MC algorithm improves over them or not. We reason that taking the algorithms as they come “out of the box” is what the vast majority of researchers who are not experts on event-detection algorithms do when choosing algorithms for purposes of data analysis.

As previous research showed that increased noise amplitude may affect the number of fixations and consequently fixation duration (Holmqvist et al., 2012; Wass et al., 2013), we calculated the number of fixations, mean fixation duration, and standard deviation of fixation duration for all noise levels and data loss levels. Such an approach is similar to Zemblys & Holmqvist (2016), who investigated the parameter settings of event-detection algorithms as a function of increasing noise level. However, their aim was to see how the settings of an algorithm should be adapted for the output to approach that of their golden standard (the algorithm they consider to be the “best”). Here, however, we compare each algorithm against itself to examine robustness to noise and data loss level of the outcome measures. After examining the noise-robustness of the outcome measures of the I2MC algorithm in this manner, we examine the application of the I2MC algorithm to infant data.

## 4.4. Algorithm

The I2MC algorithm is comprised of three separate steps: interpolation of missing data, 2-means clustering (i.e. the selection of fixation-candidates by the search rule), and finally fixation labeling (the categorization rules). It is important to note that while example values will be provided with the algorithm for all parameters, along with a motivation for the specific value, they may need to be adapted to better suit a specific data set. A flow-chart for the algorithm is depicted in panel A of Figure 4.2. A MATLAB implementation of the I2MC algorithm is freely available from [dx.doi.org/10.5281/zenodo.59249](https://dx.doi.org/10.5281/zenodo.59249)

### 4.4.1. Interpolation of missing data

To maximize the amount of eye-tracking data that can be used for event detection, imputation of short periods of missing data is performed through interpolation. We chose an interpolation method satisfying two conditions: 1) interpolation must be locally determined by the gaze samples at each end of the interpolation window and 2) interpolation must be monotonic (i.e. there are no extrema in the data points between the start and end points). The interpolation method adopted here, which satisfies both constraints, is by Steffen (1990). It should be noted that commonly used cubic spline interpolation (e.g. Frank, Vul, & Johnson, 2009), does not satisfy the constraints posed here as it can produce extrema in the interpolated data.

Interpolation was performed as follows. Periods of missing coordinates in the gaze-coordinate signals are interpolated provided that the following criteria are met. First, the period of missing coordinates has to be shorter than a set value. The value used in this paper was 100 ms, and was chosen so as not to interpolate over entire saccade-fixation-saccade sequences (saccades with a latency of 100 ms are considered extremely early; Fischer & Ramsperger, 1984). In addition, blink durations are usually higher than 100 ms, and so relatively few blinks should be interpolated. The value of 100 ms is, however, not fixed, and may be adapted according to the periods

#### 4. Noise-robust fixation detection

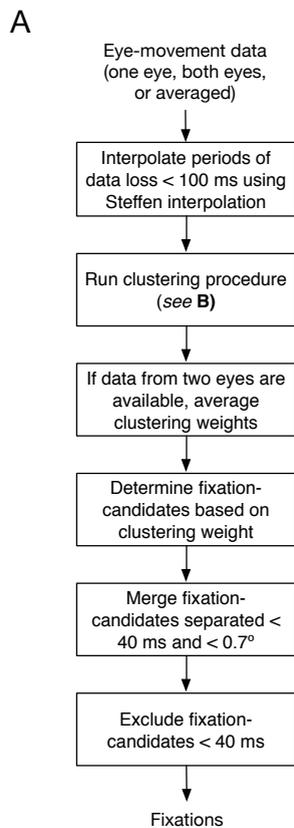
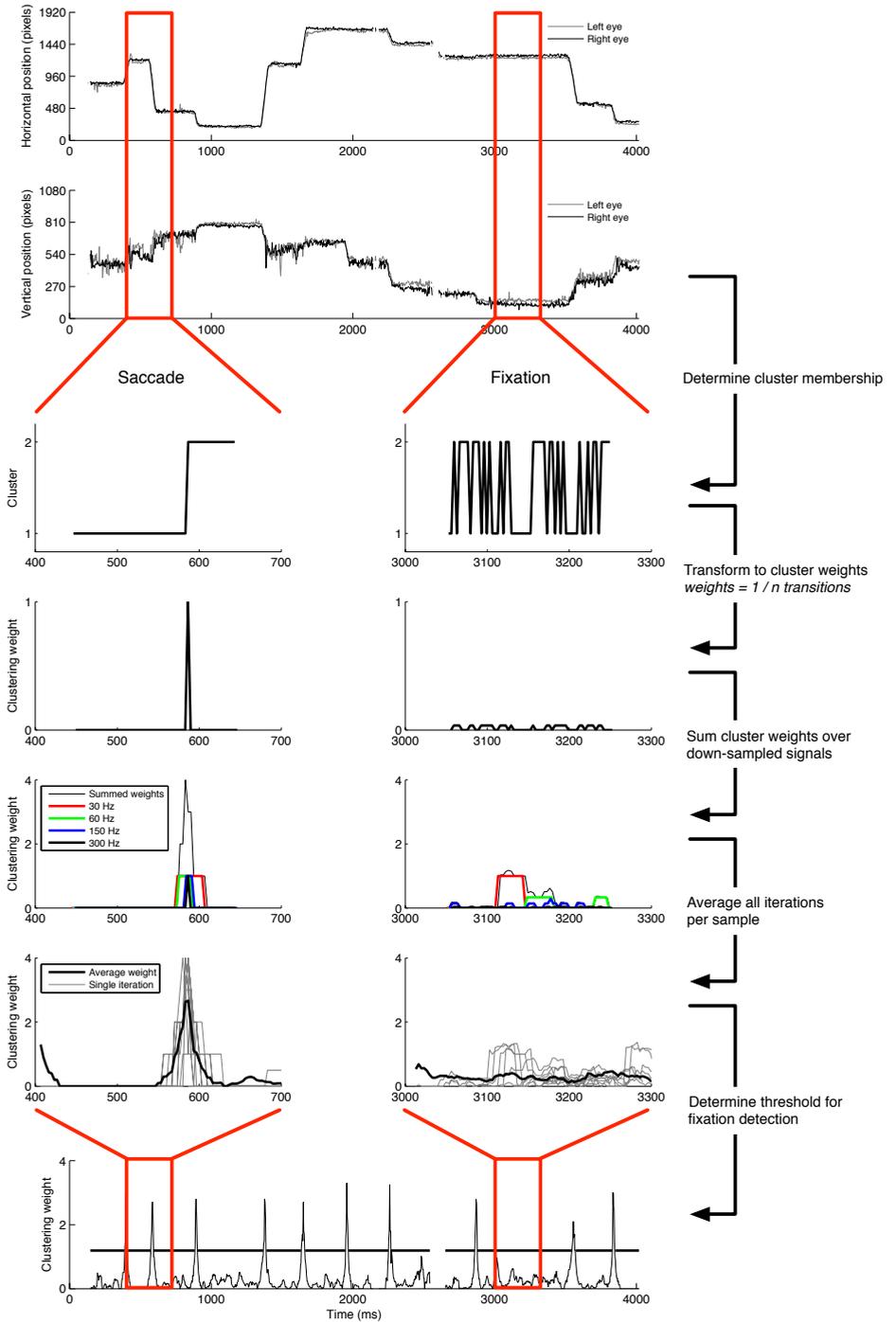


Figure 4.2.: Overview of I2MC algorithm. (A) flow-chart for the entire algorithm and (B) the specific steps of the clustering procedure (as outlined in 2-Means Clustering).

B



#### 4. Noise-robust fixation detection

of data loss observed in the eye-tracking data. Second, valid data for at least 2 samples has to be available at each end of the missing window.

##### 4.4.2. 2-Means Clustering

Following interpolation, a moving window of 200 ms width slides over the gaze position signal. The value of 200 ms was chosen so that a window generally contains at most parts of 2, not more, fixations. For each window, a 2-means clustering procedure is carried out. 2-means clustering is a variant of  $k$ -means clustering (where  $k = 2$  in this case), a procedure in which a number of observations are clustered iteratively into  $k$  clusters (see e.g. Jain, 2010). The observations belonging to each cluster are those that are closer to the mean of that cluster's observations than to the mean of any other cluster. In the present application, portions of the gaze position signal within a moving window are forced into 2 clusters. The overarching idea is that if the gaze position signal in a given window contains a saccade, there will be few cluster membership-transitions and these will be concentrated around a specific point in time – the time point of the saccade. If, however, the gaze position signal in a given window only contains a fixation, the cluster membership-transitions between the two clusters are driven only by the noise in the fixations. They may thus occur frequently and are likely spread out across the whole window. The specific algorithmic steps of the clustering procedure, which are depicted in Figure 4.2 panel B, are as follows:

1. If the current window contains no missing data (go to step 4 if it does): force the gaze position signal into two clusters. Cluster membership is a value of either 1 or 2 (i.e. the cluster the sample belongs to). It is important to note that cluster membership itself is not relevant, but only where the membership transitions from cluster 1 to cluster 2, or vice versa, occur. Only the times at which these transitions from one cluster to another occurs are used in the next step.
2. Construct a clustering weight for the current window from the cluster-membership determined in step 1. The clustering weight for samples

where a cluster-membership transition occurs is  $1 / \text{number of total transitions in the window}$ . The clustering weight for the other samples (i.e. those at which no transition occurs) is 0. If, for example, one transition occurs from cluster 1 to cluster 2 as in the saccade example in Figure 4.2, the clustering weight for the sample containing the transition is 1. For all samples containing a transition in the fixation example in Figure 4.2, the clustering weight is much lower as there are many transitions from one cluster to the other in the window.

3. In order to ensure that transitions are not caused solely by high-frequency noise in the data, down-sample the gaze position signal to integer divisions of the original sampling frequency and repeat steps 1 and 2 for each down-sampled position signal. For example, the data in Figure 4.2 (and in this paper) are recorded at 300 Hz, and down sampled to 150 Hz, 60 Hz, and 30 Hz. The clustering weights for the gaze position signal at its original sampling frequency, as well as the down sampled signals, are subsequently summed.
4. Move the window in the gaze position signal. The window may be moved one sample, or a number of samples. We use a step of 20 ms (6 samples at 300Hz) here, as it provides nearly identical result to moving the window one sample but decreases computation time six-fold. This step size is, however, configurable. If the subsequent window contains missing data or is moved past the end of the data, go back in steps of one sample to determine the last possible window. If no additional windows are possible backward in time up to the previous possible window, find the first possible window after the period of missing data. As long as the end of the window does not reach the end of the gaze position signal, return to step 1.
5. For each sample, average the clustering weights assigned in steps 1-3 for each time that sample was included in the moving window. For example, if a sample was included in three windows, it will have been assigned three clustering weights, and these three weights are averaged. The subsequent clustering weight signal (see Figure 4.2)

#### 4. *Noise-robust fixation detection*

can now be used for fixation detection.

For binocular eye-tracking data, the clustering procedure described above was run on the data for the left and the right eyes separately. These two clustering weight signals were then averaged to determine the final clustering weight signal. Doing this has the advantage that if only one eye moved according to the eye tracker, which is most likely noise, it is unlikely to lead to a large peak in the clustering weight signal. For monocular eye-tracking data, and when data from only one eye was available in binocular eye-tracking data due to data loss, the clustering weight signal from that one eye is used.

##### **4.4.3. Fixation labeling**

The categorization rules for the present algorithm are as follows. A cutoff is used to determine fixation-candidates from the clustering weight signal (see panel B in Figure 4.2). Here we use a cutoff of the mean clustering weight plus 2 standard deviations. However, different cutoffs may be required for different datasets. All periods of clustering-weight signal below this cutoff are labeled as fixation-candidates. Hereafter, consecutive fixation-candidates are merged. Finally, short fixation-candidates are excluded from the output. The settings for merging fixation-candidates may depend on the stimuli used in the experiment, the noise level in the eye-tracking data, or the size of saccades of interest. Here we opt for merging fixation-candidates that are less than  $0.7^\circ$  apart and that are separated by less than 40 ms. Fixation-candidates shorter than 40 ms are removed.

The options that may be set in the algorithm and their suggested values are summarized in Table 4.1.

Table 4.1.: Settings in the I2MC algorithm and their suggested values for data at 300Hz

Setting	Used value(s) at 300 Hz	Impact when changed
Interpolation window	100 ms	Increase will lead to interpolation of blinks, decrease will lead to less periods of data loss being interpolated
Interpolation edge	6.7 ms (2 samples)	Increase will require more data points at data loss edge, and will not interpolate in the event of flicker (i.e. repetition of short period of data loss and data points). At least two samples are required.
Clustering window size	200 ms	Increase will lead to clustering procedure being more readily carried out over saccade-fixation-saccade sequence
Downsampling	150, 60, & 30 Hz	Removal of downsampling steps will lead to more susceptibility to short bursts of noise
Window step size	20 ms	We observed no difference between 3.3 ms and 20 ms
Clustering weight cutoff	2 sd above the mean	Increase will lead to fewer fixation-candidates (more conservative), decrease to more fixation-candidates (more liberal)
Merge fixation distance	0.7°	Increase will lead to more fixation-candidates being merged
Merge fixation time	40 ms	Increase will lead to more fixation-candidates being merged
Min. fixation duration	40 ms	Increase will lead to more short fixation-candidates being excluded

## 4.5. Methods

The algorithms were compared on the following outcome measures: the number of fixations, mean fixation duration and standard deviation of fixation duration. These outcome measures were obtained from eye-movement data with increasing noise amplitude and periods of data loss, as well as the combination of the two. A data set with binocular data recorded with the SR Research Eyelink 1000 by Hooge, Nyström, Cornelissen, & Holmqvist (2015) was used in which the noise amplitude or data loss was artificially increased. This data set was chosen based on the low noise amplitude, and low data loss levels. In their experiment, participants made horizontal and vertical saccades of a wide range of amplitudes.

To examine robustness, noise amplitude and data loss level in the data were artificially increased. To provide a valid test of algorithm performance, it is important that the added noise and data loss are representative of that which occurs when doing eye-tracking research in sub-optimal conditions. To achieve this, we first characterized what the noise and data loss looked like in a set of infant data recorded by Hessels et al. (2015a); see for an example the bottom panel of Figure 4.1. We subsequently developed methods to add noise and data loss with these characteristics at varying levels to our clean data. These methods are described in detail in Appendix A. Noise level was varied from a sample-to-sample RMS noise level of 0 to  $5.57^\circ$ . This latter value was chosen to be beyond the upper limit for noise typically encountered in eye-movement data. For comparison; RMS noise level in infant eye-movement data rarely exceeded  $3^\circ$  (Hessels, Kemner, van den Boomen, & Hooge, 2016b). Data loss was varied by changing the occurrence of periods of data loss from 0-100% of the trial. As noise and data loss were characterized in a data set recorded at 300 Hz, our clean data to which we subsequently added noise and data loss was down sampled from 1000 Hz to 300 Hz using 1st order interpolation.

After examining the noise-robustness of I2MC and competing algorithms in eye-movement data with a range of artificially generated noise and data

loss levels, we apply the algorithms to infant data. In order to interpret the outcome measures of the algorithms when applied to infant data, four eye-movement experts (authors RH, DN and IH, and an external expert) hand coded fixations in the infant data. A subset of infant data was extracted from Hessels, Hooge, & Kemner (2016a) for manual coding. Data from 20 infants amounting to a total of 40 minutes of eye-movement data was coded. Hand coding was done in custom MATLAB software and took each manual coder approximately 3 hours to complete.

#### 4.5.1. Algorithms for comparison

Seven algorithms were chosen for comparison against the I2MC algorithm. As only 3 of the 7 algorithms provided output for all noise and/or data loss levels, only these algorithms are discussed here. The remaining four algorithms are included in Appendix B. For each algorithm, only parameters that are dependent on sample frequency of the input data were adjusted. They were set to match their initial values for the 300 Hz data we use here. The search and categorization rules, as well as the motivation for including each algorithm, are described below.

**Adaptive velocity algorithms for low-frequency data (HC).** An implementation of an adaptive velocity search rule is given by Hooge & Camps (2013). Their algorithm labels fixations instead of saccades and was originally designed for low-frequency (120 Hz or lower) eye-movement data. The categorization rules that they employ are: 1) adjacent fixations that are less than  $1.0^\circ$  away from each other are merged, and 2) fixations shorter than 60 ms are excluded. The reason for this algorithm’s inclusion is that it’s a simple fixation-labeling algorithm with few parameters, that uses a threshold that is adaptive to the noise level in the data.

**Binocular-Individual Threshold (BIT).** Another implementation of an adaptive velocity search rule is given by van der Lans, Wedel, & Pieters (2011). They describe their algorithm as follows: “*Our Binocular-Individual Threshold (BIT) algorithm for identifying fixations is [...] a parameter-*

#### 4. Noise-robust fixation detection

*free fixation-identification algorithm that automatically identifies task- and individual-specific velocity thresholds by optimally exploiting the statistical properties of the eye-movement data across different eyes and directions of eye movements”* (p. 240). The algorithm improves over standard adaptive velocity search rules by using the covariance between movement of the left and right eye. If the left eye moves in a given direction, the right eye often does so too, whereas for noise the movement of the two eyes is uncorrelated. Given this feature, the BIT algorithm may more readily be able to distinguish saccades from noise compared to standard velocity algorithms, and it is therefore included in this comparison. No further categorization rules are reported by the authors of the algorithm.

**Identification by analysis of variance and covariance (CDT).** Veneri et al. (2011) designed a fixation-labeling algorithm, with a search rule based on the covariance of the horizontal and vertical eye position signals. The algorithm labels gaze samples as belonging to a fixation when an F-test indicates that the variance of the horizontal and vertical eye positions is equal. When covariance between x- and y-coordinates is high, samples are labeled as belonging to a saccade. The remaining samples are labeled according to a combination of their covariance and the F-test for equal variance. Veneri et al. (2011) report that their CDT algorithm more accurately identified fixations compared to a standard dispersion algorithm when the noise amplitude was high. The reason for its inclusion is that it appears to be robust to noise. No further categorization rules are reported by the authors of the algorithm.

## 4.6. Results

### 4.6.1. RMS noise

The number of fixations, mean fixation duration and standard deviation of fixation duration were calculated for all algorithms as a function of the RMS noise level added to the eye-movement data. As depicted in Figure 4.3, there is a lot of variation in how the algorithms’ output was affected by

the noise level. The HC and BIT algorithms showed an immediate decrease in the number of fixations detected as noise increased. The number of fixations for the HC algorithm slowly decreased but did not reach zero. For the BIT algorithm, the number of fixations stabilized for RMS noise larger than  $2^\circ$ . The CDT algorithm showed a steady increase in the number of fixations detected as noise level increased. Finally, the I2MC algorithm produced a fairly consistent number of fixations as a function of noise level, with only a small increase in the total number when more than  $5^\circ$  of RMS noise was added to the eye-movement data.

The HC algorithm showed an increase in mean fixation duration as noise increased and an increase in the standard deviation of fixation duration. The BIT algorithm showed an increase in both mean fixation duration and standard deviation of fixation duration. The CDT algorithm showed a slowly decreasing, but still fairly consistent mean fixation duration and standard deviation of fixation duration. Finally, the I2MC algorithm showed a fairly consistent mean fixation duration and standard deviation of fixation duration as a function of noise level, with only a small decrease in mean fixation duration when more than  $5^\circ$  of RMS noise was added to the eye-movement data.

These findings show that the outcome measures from the I2MC algorithm were most robust among the algorithms we tested to increasing noise level in the eye-movement data.

#### 4.6.2. Variable RMS noise

In order to determine how robust algorithms were to large variations in noise level during in a single portion of the eye-movement data, RMS noise was level was increased in segments totalling half of the trial. This mimics short bursts of noise as may for instance occur when tracking is unstable for one part of the screen. The eye-movement data in each trial thus contained both periods with and without added RMS noise. As depicted in Figure 4.4, there was again some variation between the algorithms' reported number

#### 4. Noise-robust fixation detection

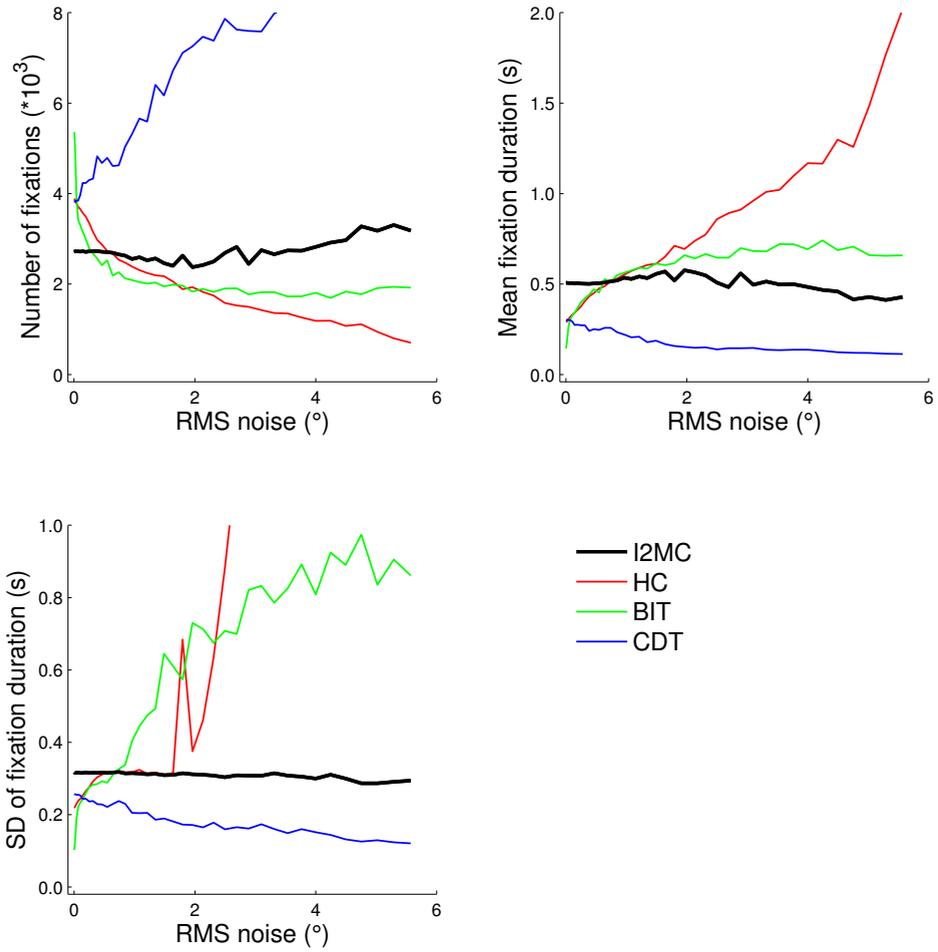


Figure 4.3.: Number of fixations (top left panel), mean fixation duration (top right panel), and standard deviation of fixation duration (bottom left panel) for four event-detection algorithms as a function of RMS noise added to the eye-movement data.

of fixations, mean fixation duration, and standard deviation of fixation duration. The results for the number of fixations detected closely resembled the RMS noise analyses, although decreases and increases were smaller. The BIT and HC algorithms showed a decrease in the number of fixations. The CDT algorithm showed an increase in the number of fixations. Finally, the I2MC algorithm showed a stable number of fixations as a function of variable RMS noise level.

For the mean fixation duration and standard deviation of fixation duration the result grossly matched the previous RMS noise analyses, albeit with smaller in- and decreases.

### 4.6.3. Data loss

The number of fixations, mean fixation duration and standard deviation of fixation duration were calculated for all algorithms as a function of the amount of data loss added to the eye-movement data, from 0% (no data loss occurs) to 100% (data loss can occur throughout entire trial). As depicted in Figure 4.5, the difference between algorithms was much smaller than for the previous analyses. The CDT algorithms showed an increase in the number of fixations as data loss increases. The BIT algorithm, on the other hand, showed a decrease in the number of fixations as data loss increased. Both the HC and I2MC algorithm were stable in the number of fixations as a function of data loss added to the eye-movement data.

For the CDT algorithm, the mean fixation duration and standard deviation of fixation duration decreased as a function of data loss. For the HC algorithm, there was a pattern of a slight decrease, followed by an increase, and then again a decrease for both mean fixation duration and standard deviation of fixation duration as a function of data loss. Finally, for the BIT and I2MC algorithms, the mean fixation duration was fairly stable as data loss increased.

#### 4. Noise-robust fixation detection

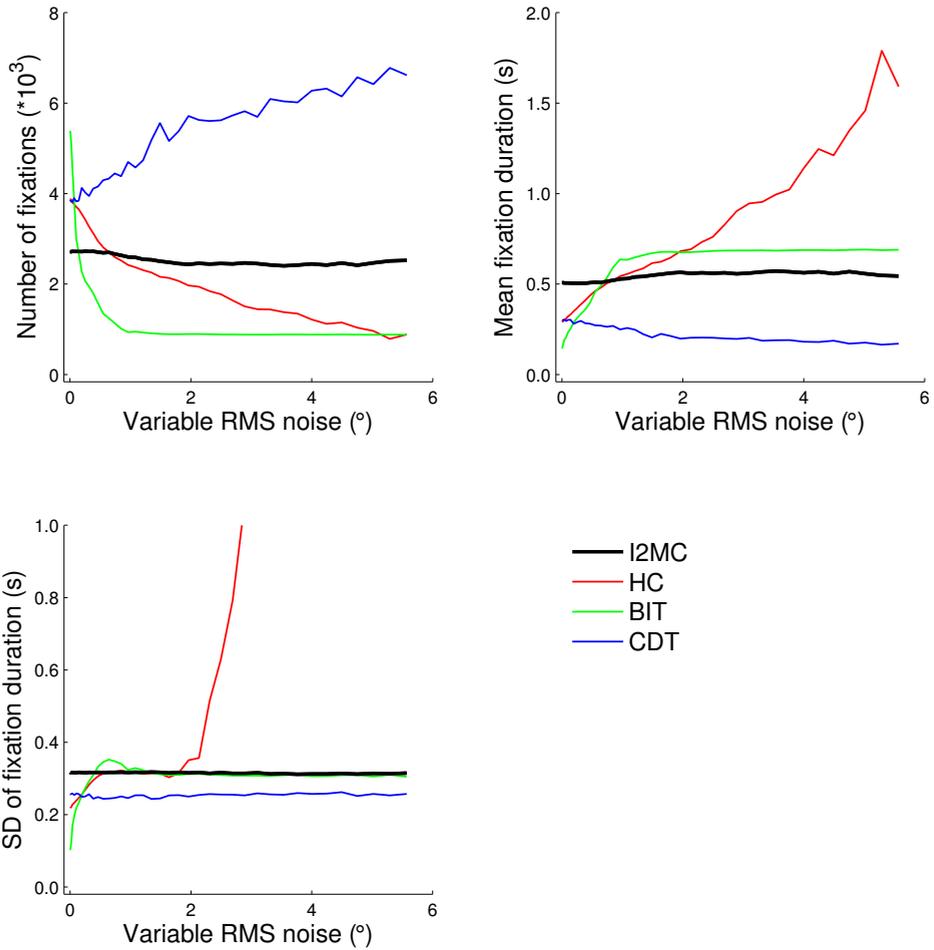


Figure 4.4.: Number of fixations (top left panel), mean fixation duration (top right panel), and standard deviation of fixation duration (bottom left panel) for four event-detection algorithms as a function of variable RMS noise added to the eye-movement data.

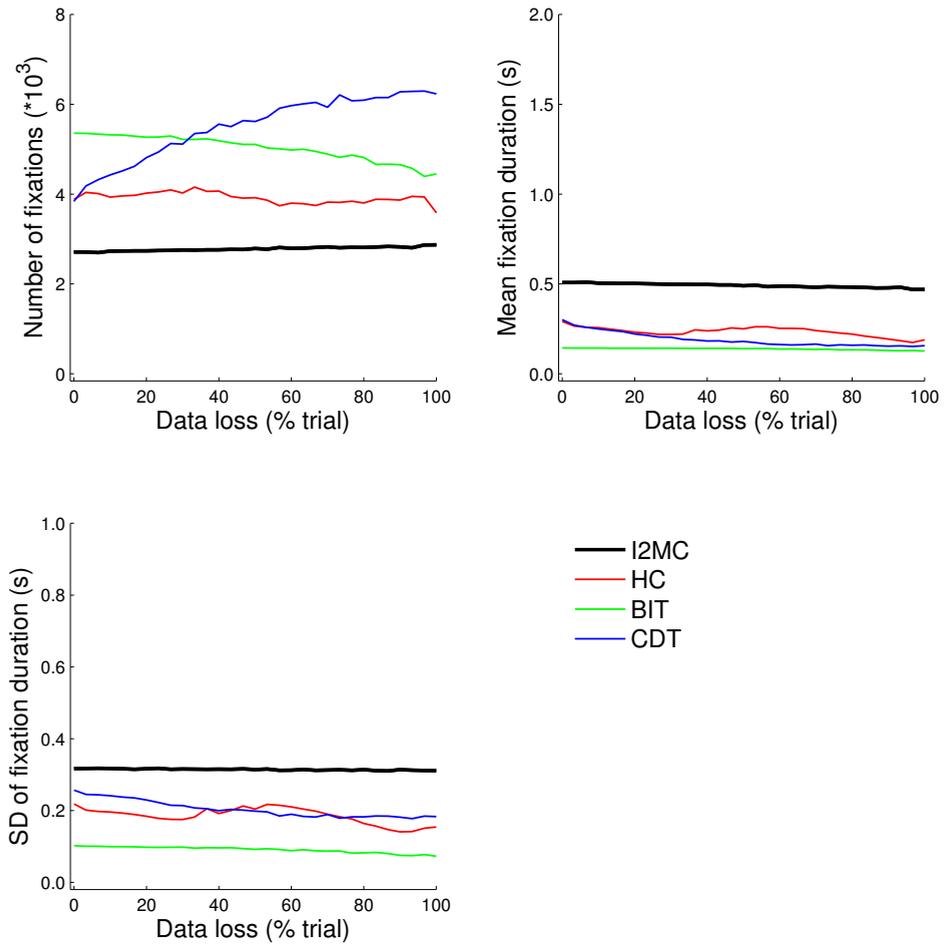


Figure 4.5.: Number of fixations (top left panel), mean fixation duration (top right panel), and standard deviation of fixation duration (bottom left panel) for four event-detection algorithms as a function of data loss added to the eye-movement data.

#### 4. Noise-robust fixation detection

##### 4.6.4. Combined noise and data loss

As a final test of the robustness of the outcome measures of the four algorithms, a combination of RMS noise and data loss was added to the eye-movement data.

As described in the noise analysis, and visible from the left column in Figure 4.6, there was only a small increase in the number of fixations detected by the I2MC algorithm as RMS noise increased. When both data loss and noise increased, the number of fixations detected was somewhat larger than when only noise level increased. Consequently, the mean fixation duration and standard deviation of fixation duration decreased with increasing noise level, and more so when data loss also increased. For the three competing algorithms (HC, BIT, and CDT), the differences were markedly larger. As visible from the second column in Figure 4.6, the number of fixations decreased for the HC algorithms as noise level increased. Moreover, the differences between the levels of data loss were large, and the number of detected fixations increased four-fold as data loss increased at the highest noise level. For the HC algorithm, the mean fixation duration and standard deviation of fixation duration as a function of noise level increased most for the lowest data loss level, and appeared to be most robust for the highest data loss level (we return to this apparent robustness shortly). As visible from the third column in Figure 4.6, the results for the BIT algorithm are similar to the results from the HC algorithm. However, while the HC algorithm showed an increasing number of fixations, and decreasing mean fixation duration and standard deviation of fixation duration across all noise levels, the values for the BIT algorithm appeared to stabilize for RMS noise levels larger than  $2^\circ$ , albeit with marked differences between the levels of data loss. Finally, as visible from the right column in Figure 4.6, the number of detected fixations for the CDT algorithm increased as a function of noise level for all levels of data loss. Consequently, the mean fixation duration and standard deviation of fixation duration decreased as a function of noise level for all levels of data loss. Concluding, the number of fixations, mean fixation duration, and standard deviation of fixation du-

ration were most robust for the I2MC algorithm as noise level and/or data loss increased.

#### 4.6.5. Interpreting noise-robustness

For three of the algorithms there was at least one data loss level where the outcome measures were robust to changes in noise level. First, the outcome measures from the I2MC algorithm were most robust of all the algorithms to both noise level and data loss level and the combination thereof. Second, the outcome measures for the BIT algorithm were robust to increases in noise level, albeit with marked differences between the different levels of data loss. Third, the outcome measures reported by the HC algorithm were robust to increases in noise level for the highest data loss level. In order to interpret this noise robustness we examined the precise differences in the distributions of fixation durations between these algorithms. 2D histograms of fixation duration were computed for the varying noise and data loss levels. As visible from Figure 4.7 (top panels), the distribution of fixation durations detected by the I2MC algorithm remains almost unchanged from 0-0.85° RMS noise and 0-100% data loss. Note that, as there are a lot of fast corrective saccades following undershoot of the target in this data set, there is a large peak in fixation duration around 140 ms. For higher RMS values, fewer short fixations are reported, and consequently relatively more longer fixations are reported. The differences in the distribution of fixation durations as a function of data loss remain minimal for the 1.96° RMS noise level. For the higher RMS noise levels (3.53° and 5.57°), a larger number of longer fixation durations are reported in general, compared to lower RMS noise levels. Moreover, more short fixations are reported for the higher data loss levels for these two RMS noise levels.

For both the HC and BIT algorithms (middle two rows of Figure 4.7) there are few differences in the distribution of fixation durations as data loss increased for the lowest RMS noise level. For an RMS noise level of 0.85° and up, however, both algorithms report progressively less short fixation durations for the lowest data loss level. When data loss increased

#### 4. Noise-robust fixation detection

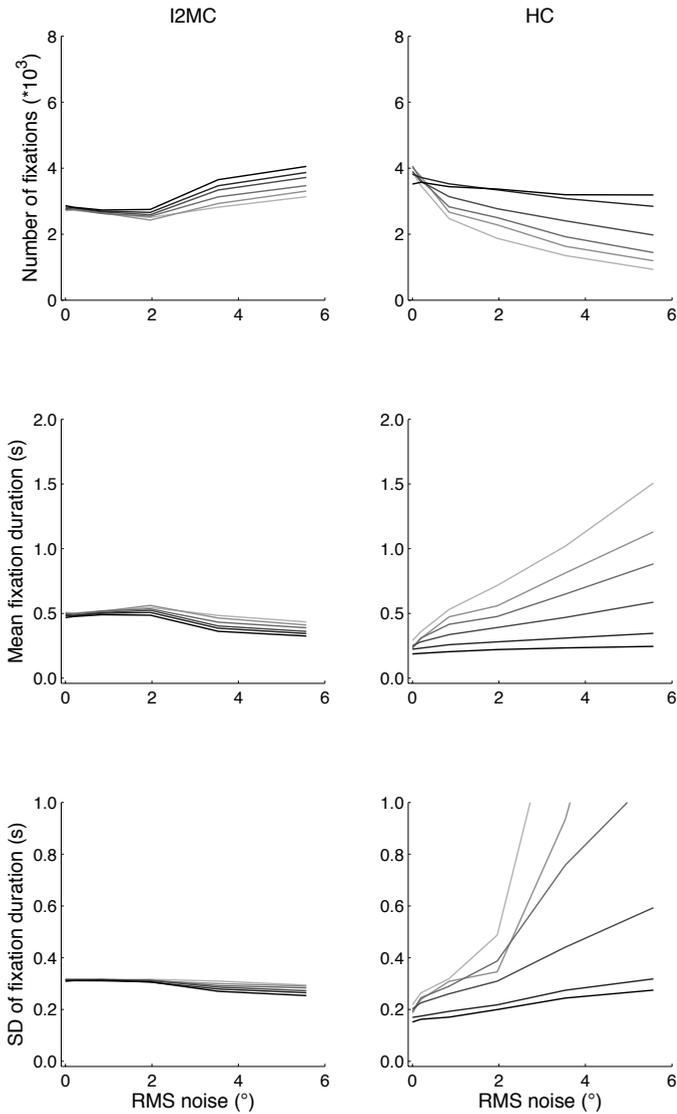
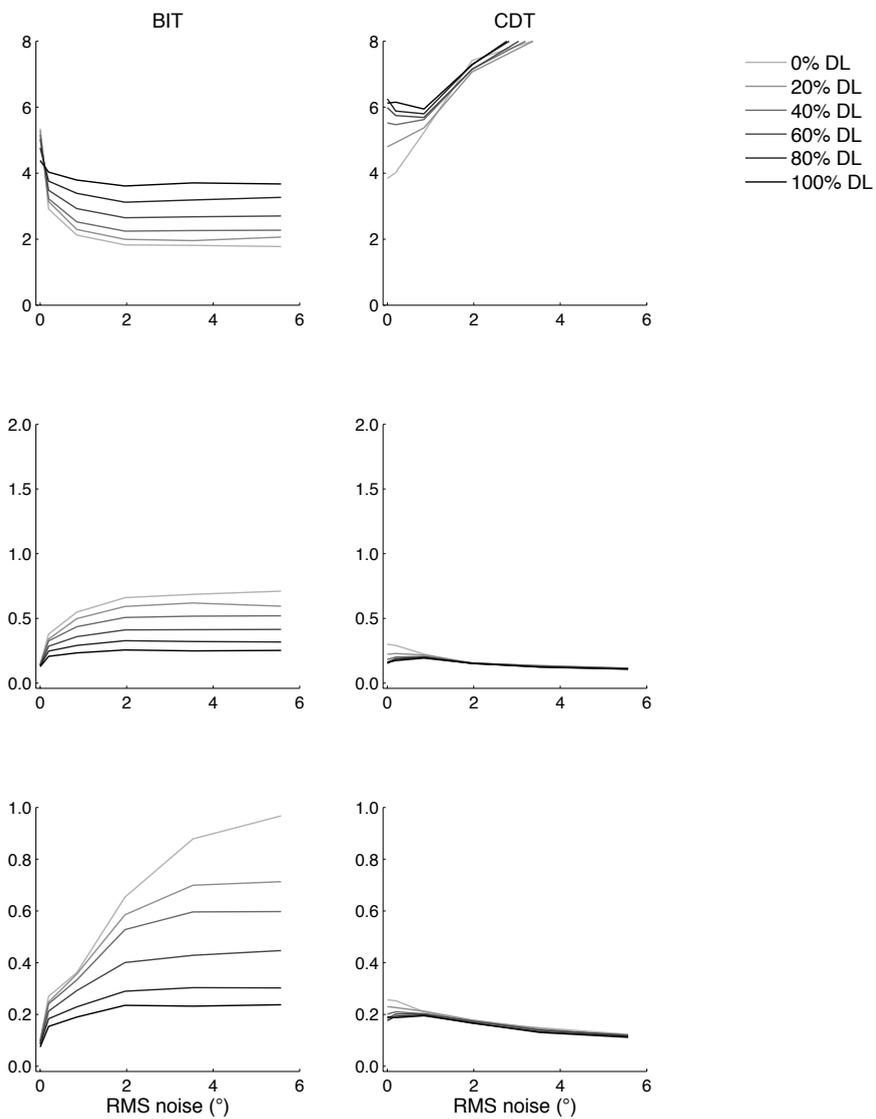


Figure 4.6.: Number of fixations (top panels), mean fixation duration (middle panels), and standard deviation of fixation duration (bottom panels) for the final four algorithms as a function of noise level in the eye-movement data. From left to right columns depict the I2MC, HC, BIT, and CDT algorithms. Separate lines indicate data loss added to 0% (lightest grey) to 100% (black) of trial.

## 4.6. Results



#### 4. Noise-robust fixation detection

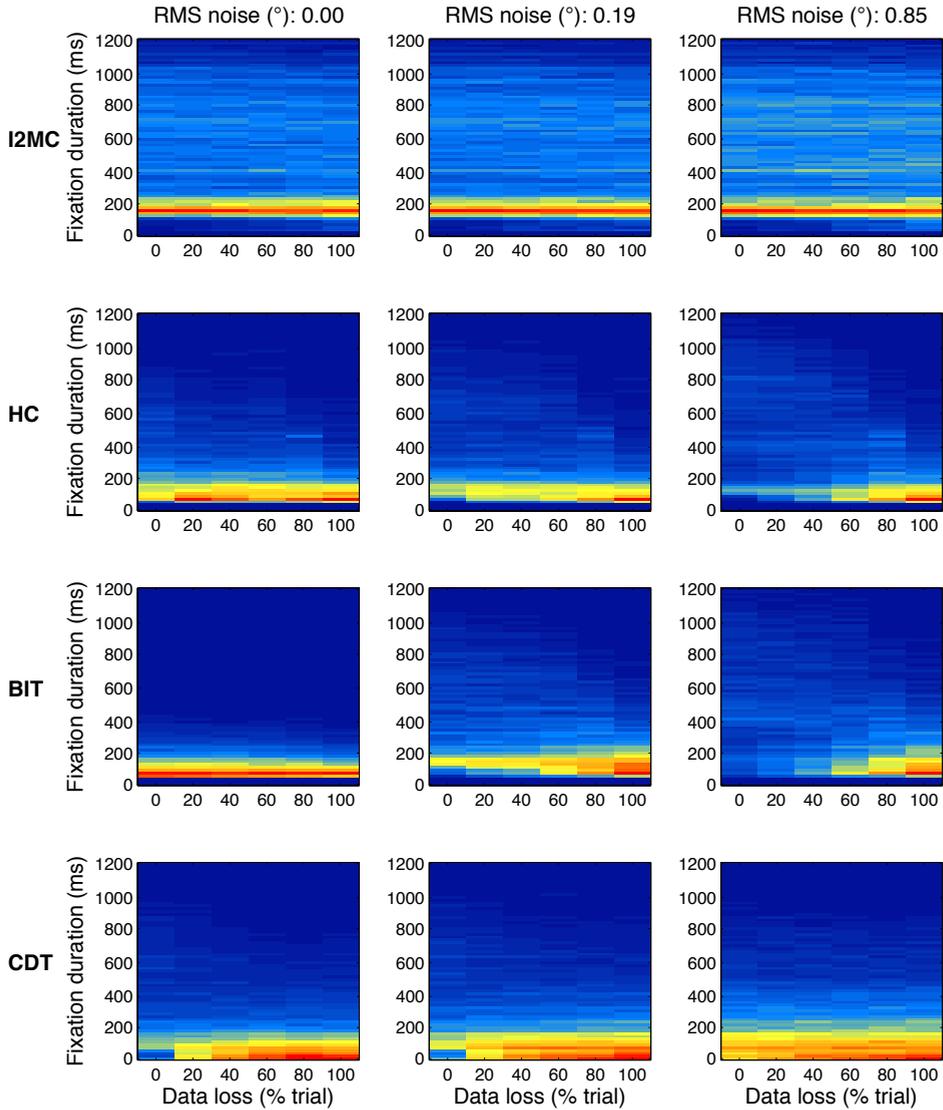
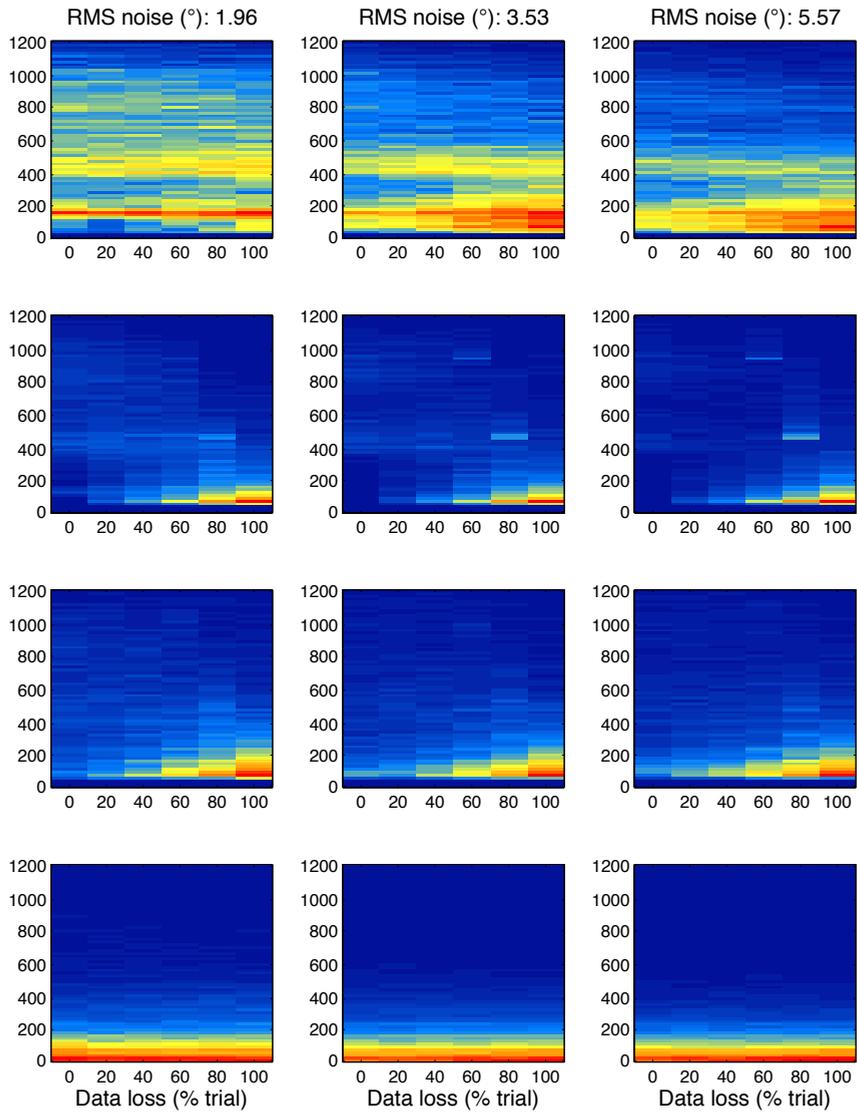


Figure 4.7.: 2D histograms of fixation duration for fixations detected by the I2MC, HC, BIT, and CDT algorithms. Columns depict different noise levels from low (left) to high (right) noise. Within each histogram 5 levels of data loss are depicted from 0% (left) to 100% (right) of the trial. Redder colors indicate more detected fixations, bluer colors less detected fixations.



#### 4. Noise-robust fixation detection

progressively more short fixations are again reported. This suggests that both the HC and BIT algorithm detected longer fixations when noise increased, which were subsequently broken up into shorter fixations when data loss increased. Finally, for the CDT algorithm, progressively more short fixations were reported when RMS noise exceeded  $1.96^\circ$ . The differences between the distributions of fixation durations for the varying data loss levels were minimal, most probably due to the fixation durations already being extremely short. As visible from Figure 4.8, the number of fixations and fixation duration detected by the I2MC algorithm were most noise-invariant as compared to the other three algorithms. Concluding, the number of fixations and the distribution of fixation durations in the output from the I2MC algorithm was affected the least by noise level and data loss level.

#### 4.6.6. Applying the algorithm to infant data

After ascertaining the noise-robustness of the I2MC algorithm, it and the competing algorithms were applied to 40 minutes of infant data. The average number of fixations per trial and the mean fixation duration were calculated for the I2MC, HC, BIT, and CDT algorithms, as well as for the four expert coders. The algorithm that best approaches the average of the four expert coders is considered best in the application to infant data. As visible from Figure 4.9, there was some variability in the outcome measures of the expert coders. Moreover, the expert coders generally detected fewer fixations per trials, and the mean fixation duration was generally longer as compared to the algorithms. When comparing the algorithms to the average of the four expert coders, a couple of conclusions can be drawn. The HC and BIT algorithms detected approximately 1-2 fixations more per trial than the average of the expert coders, and the mean fixation duration was between 100-150 ms shorter than the average of the expert coders. Both the CDT and I2MC algorithm better approximated the average of the expert coders, and the I2MC did so best. The difference between the I2MC algorithm and the average of the four expert coders was 0.77 fixations per trial, and 24 ms in mean fixation duration.

Three representative excerpts of infant eye-movement data with varying noise levels are depicted in Figure 4.10. When considering the trial with the lowest noise level (left panel), the number and duration of the fixations detected by the four algorithms and four coders are highly similar. However, for the trial with higher noise levels (middle panel), HC and BIT detected more fixations than I2MC, CDT, and expert coders. Finally, for the trial with the highest noise level (right panel), CDT also detected more fixations than the four expert coders. Across the different noise levels, the I2MC algorithm best agreed with the four expert coders. The main difference between the I2MC algorithm and the expert coders is that the fixation durations are slightly longer for the I2MC algorithm. In conclusion, the I2MC algorithm not only performed best in eye-movement data with artificially increased noise and data loss, but also when applied to actual infant data.

## 4.7. Discussion

The purpose of the present work was to address the need for an algorithm capable of labeling fixations across a wide range of noise levels in data where periods of data loss may occur. This is particularly relevant given the rise of remote video eye tracking in participant groups such as infants, school children, and certain patient groups where body movement is difficult to restrain and may strongly affect eye-movement data quality. Here we proposed and evaluated a new algorithm designed specifically for eye-movement data of varying data quality: Identification by 2-Mean Clustering (I2MC).

In comparison with other state-of-the-art event-detection algorithms, we found the following. First, we report that the number of fixations, mean fixation duration and standard deviation of fixation duration by the I2MC algorithm were most robust to increases in noise level, compared to the competing algorithms. This was the case both when the noise level was increased for the entire eye-movement signal, and when the noise level was increased for only part of the eye-movement signal, mimicking short bursts of noise. Second, differences between algorithms were smaller for the data

#### 4. Noise-robust fixation detection

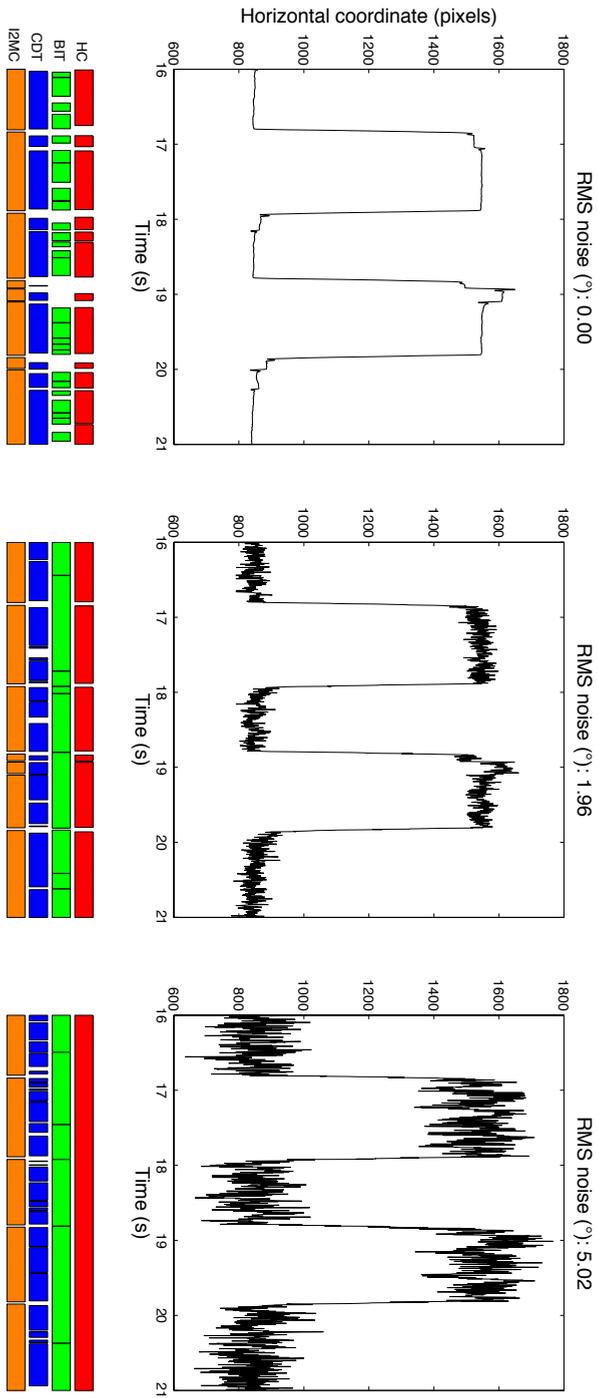


Figure 4.8.: Fixations labeled by the four algorithms for data prior to adding RMS noise (left), after  $1.96^\circ$  (middle), and  $5.02^\circ$  (right) of added RMS noise. Fixations labeled by the algorithms are presented in the bottom panels. Data from a trial in which  $10^\circ$  horizontal saccades had to be made.

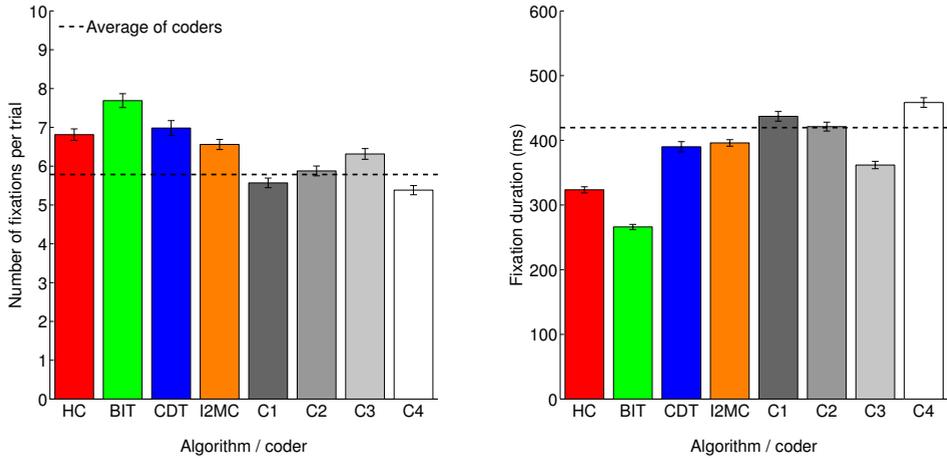


Figure 4.9.: Average number of fixations per trial (left) and average fixation duration (right) for the infant eye-movement data as reported by the four algorithms and four coders. Error bars represent standard error of the mean. The dashed line indicates the average of the four coders.

loss analysis compared to the noise analyses. However, the outcome measures of the I2MC and BIT algorithms were most robust to increases in data loss. We report that when adding both noise and data loss to the eye-movement data, the outcome measures for the I2MC algorithm were most robust. This was particularly the case for  $0-2^\circ$  of noise where outcome measures were almost identical. As the BIT algorithm appeared to show stable outcome measures as a function of noise for noise amplitudes larger than  $2^\circ$ , albeit with marked differences between data loss levels, and the HC algorithm showed stable outcome measures for the highest data loss level only, we examined the distributions of fixation duration more closely. Here, we report that the I2MC algorithm showed nearly identical distributions of fixation durations for the lowest noise levels regardless of data loss level, whereas differences in the distribution of fixation duration for the competing algorithms were already present at the lowest noise levels. However, when noise amplitude was larger than  $2^\circ$ , the distribution of fixation duration of the I2MC algorithm also began to show marked differences

#### 4. Noise-robust fixation detection

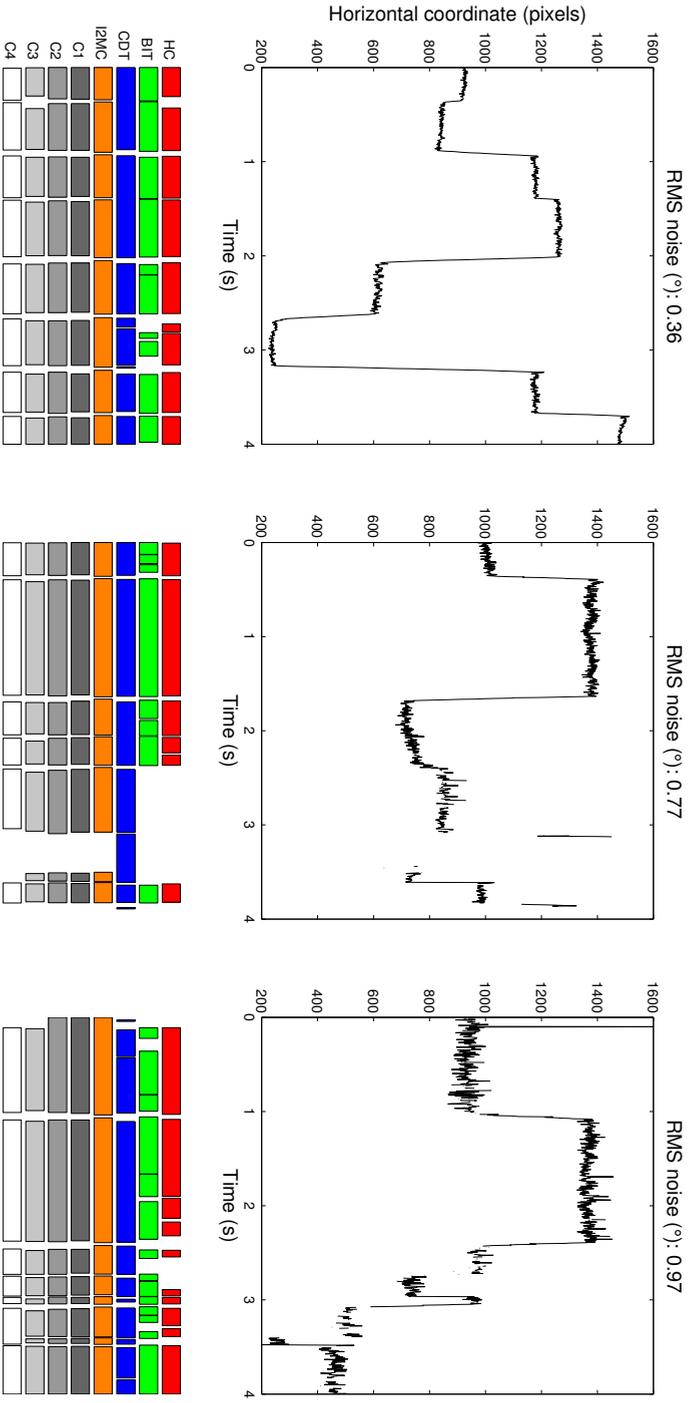


Figure 4.10.: Fixations labeled by the four algorithms and four coders for three representative trials containing infant eye-movement data of varying noise level. Fixations labeled by the algorithms and coders are presented in the bottom panels. Noise levels were estimated by computing the RMS noise of the longest fixation detected by the I2MC algorithm.

compared to the lower noise levels. Finally, when the I2MC was applied to infant data, the outcome measures best approached that of the average of four expert coders, as compared to the other algorithms. Concluding, the outcome measures for the I2MC algorithm were most robust to noise, most notable in the range 0-2° of RMS noise (for each data loss level) added to the eye-movement data.

If we compare the noise-robustness of the I2MC algorithm to previous research, we see it also compares favorably to manufacturer-provided algorithms. For instance, Zemblys & Holmqvist (2016) report that the number of fixations and mean fixation duration for the SMI I-VT and I-DT algorithms change drastically for noise levels higher than 0.25° RMS. In addition, they created a mathematical model to estimate the best-compromise threshold for the SMI dispersion and velocity-threshold algorithms from the noise level observed in the eye movement recording. While adjusting the threshold may compensate for higher noise levels, this comes at a cost of reduced agreement of the algorithms' output (i.e. number of fixations, mean fixation duration) compared to their golden standard. Moreover, a large-scale study of data quality found that many of the remote eye trackers produced data with noise levels over 2° for some of the participants, with infrequent recordings where RMS noise level was even over 3° (Zemblys & Holmqvist, 2016). In addition, Hessels et al. (2016b) report that the RMS noise is rarely over 3° in infant research. The fact that the outcome measures of the I2MC are noise-robust, and particularly so between 0° and 2° RMS noise, means that the algorithm may apply to most real situations. Moreover, no parameters need to be adjusted to achieve the same output of fixation parameters when noise level varies between 0 and 2°, whereas this is the case for manufacturer-provided algorithms in this range (Zemblys & Holmqvist, 2016).

As noted in our introduction, a fixation is defined by how it is computed. This means that a fixation for one algorithm is not the same as a fixation for the other algorithm. Even when noise level was low in the eye-movement data investigated here, there were large differences between

#### 4. *Noise-robust fixation detection*

algorithms in the number of fixations and fixation duration (see also Appendix B). This is not only a result of the search rule algorithms apply to find fixation-candidates, but also of the categorization rules used to merge or exclude these fixation-candidates. When one algorithm excludes more fixation-candidates under uncertainty, it will produce fewer fixations as output, compared to an algorithm that does not exclude any fixation-candidates. As such, comparing fixation parameters between algorithms is like comparing apples and oranges. Instead of doing so, we first compared fixation parameters for each algorithm to itself as a function of data quality – the absolute values for each algorithm are not so important, only the change in the value as a function of noise or data loss. In essence, we compared apples to apples, and oranges to oranges. Here we presented an algorithm that produces the same output under a wide range of circumstances – the apple remains roughly the same regardless of the situation. Does this then mean that the output of the presented algorithm is “good”? That is a very difficult question to answer. Commonly in the literature, this question is tackled by comparing the output of an algorithm to an expert coder. Here, when comparing the outcome measures of I2MC and competing algorithms to four expert coders, I2MC also outperformed the other algorithms. It should be noted, however, that the expert coders did not produce identical outcome measures, such that the question beckons how informative one expert coder actually is. Future research should examine whether expert coders serve as a good golden standard for event-detection algorithms.

The I2MC algorithm is applicable to fixation labeling in situations where the data quality may be low, for instance when working with infants, school children, or certain patient groups. The I2MC algorithm may also be used when the noise and data loss levels are markedly different between trials and/or subjects; the output should be comparable despite these differences in noise and data loss levels. This is also particularly relevant for studies where two groups are compared: For instance, Shic, Chawarska, & Scasellati (2008) report that changing the parameters of a fixation-detection algorithm may reverse effects between toddlers with autism spectrum disor-

der and typically developing controls. In addition, recent work also suggests that differences in data quality between groups should be carefully monitored (Keehn & Joseph, 2016). Here, using an algorithm for which the outcome measures are robust to differences in data quality between groups may be a better solution.

While the I2MC algorithm may be applicable in a wide range of studies, there are several limitations. First, the I2MC algorithm is only built for labeling fixations. When saccade parameters are of interest, for example, the I2MC is not a sensible analysis tool. Second, we restricted the present study to labeling fixations in data collected with remote or tower-mounted eye trackers using static stimuli. When head-mounted eye trackers are used, the gaze position signal is complicated by periods of optokinetic nystagmus and the vestibulo-ocular reflex, which may require a different event-detection strategy. Moreover, when moving stimuli are used in remote or tower-mounted eye trackers, smooth pursuit movements may occur. Recent work has begun to address event detection in these situations (Larsson et al., 2015). In addition, a general limitation to the present work is that we focused on 300Hz data, while there is a broad range of sampling frequencies being used in eye-tracking research. Moreover, while the selection of algorithms against which we tested the I2MC algorithm was motivated, it represents but a subset of the entire event-detection catalogue.

## 4.8. Conclusion

Here we presented a fixation-detection algorithm for eye-tracking data recorded with remote or tower-mounted eye trackers using static stimuli. The algorithm works offline and is automatic. The key improvement made by this algorithm is that it labels fixation across a wide range of noise levels and when periods of data loss may occur.

## **Acknowledgements**

The authors would like to thank Jacco van Elst for valuable help in the coding of eye-movement data. The study was financed through the Consortium on Individual Development (CID). CID is funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant number 024.001.003 awarded to author CK). The funding body had no involvement at any stage of the study.

## References

- Andersson, R., Larsson, L., Holmqvist, K., Stridh, M., & Nyström, M. (2016). One algorithm to rule them all? An evaluation and discussion of ten eye movement event-detection algorithms. *Behavior Research Methods*.
- Aslin, R. N., & McMurray, B. (2004). Automated corneal-reflection eye tracking in infancy: Methodological developments and applications to cognition. *Infancy*, 6(2):155–163.
- Coey, C. A., Wallot, S., Richardson, M. J., & van Orden, G. (2012). On the structure of measurement noise in eye-tracking. *Journal of Eye Movement Research*, 5(4):1–10.
- Engbert, R., & Kliegl, R. (2003). Microsaccades uncover the orientation of covert attention. *Vision Research*, 43(9):1035–1045.
- Findlay, J. M. (1971). Frequency analysis of human involuntary eye movement. *Kybernetik*, 6:1–8.
- Fischer, B., & Ramsperger, E. (1984). Human express saccades: extremely short reaction times of goal directed eye movements. *Experimental Brain Research*, 57:191–195.
- Frank, M. C., Vul, E., & Johnson, S. P. (2009). Development of infants' attention to faces during the first year. *Cognition*, 110(2):160–170.
- Hessels, R. S., Andersson, R., Hooge, I. T. C., Nyström, M., & Kemner, C. (2015a). Consequences of eye color, positioning, and head movement for eye-tracking data quality in infant research. *Infancy*, 20(6):601–633.
- Hessels, R. S., Cornelissen, T. H. W., Kemner, C., & Hooge, I. T. C. (2015b). Qualitative tests of remote eyetracker recovery and performance during head rotation. *Behavior Research Methods*, 47(3):848–859.
- Hessels, R. S., Hooge, I. T. C., & Kemner, C. (2016a). An in-depth look at saccadic search in infancy. *Journal of Vision*, 16(8):10.
- Hessels, R. S., Kemner, C., van den Boomen, C., & Hooge, I. T. C. (2016b). The area-of-interest problem in eyetracking research: A noise-robust solution for face and sparse stimuli. *Behavior Research Methods*, 48(4):1694–1712.
- Holmberg, N., Holmqvist, K., & Sandberg, H. (2015). Children's attention to online adverts is related to low-level saliency factors and individual level of gaze control. *Journal of Eye Movement Research*, 8(2):1–10.
- Holmqvist, K., Nyström, M., & Mulvey, F. (2012). Eye tracker data quality: What it is and how to measure it. *Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '12*, 45.
- Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van

#### 4. Noise-robust fixation detection

- de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press.
- Hooge, I., & Camps, G. (2013). Scan path entropy and arrow plots: capturing scanning behavior of multiple observers. *Frontiers in Psychology*, 4:996.
- Hooge, I., Nyström, M., Cornelissen, T., & Holmqvist, K. (2015). The art of braking: Post saccadic oscillations in the eye tracker signal decrease with increasing saccade size. *Vision Research*, 112:55–67.
- Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8):651–666.
- Keehn, B., & Joseph, R. M. (2016). Exploring what’s missing: What do target absent trials reveal about autism search superiority? *Journal of Autism and Developmental Disorders*.
- Komogortsev, O. V., & Khan, J. I. (2009). Eye movement prediction by oculomotor plant Kalman filter with brainstem control. *Journal of Control Theory and Applications*, 7(1):1–13.
- Komogortsev, O. V., Gobert, D. V., Jayarathna, S., Koh, D. H., & Gowda, S. M. (2010). Standardization of automated analyses of oculomotor fixation and saccadic behaviors. *IEEE Transactions on Biomedical Engineering*, 57(11):2635–2645.
- Larsson, L., Nyström, M., Andersson, R., & Stridh, M. (2015). Detection of fixations and smooth pursuit movements in high-speed eye-tracking data. *Biomedical Signal Processing and Control*, 18:145–152.
- Nyström, M., & Holmqvist, K. (2010). An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. *Behavior Research Methods*, 42(1), 188–204.
- Nyström, M., Hooge, I. T. C., & Holmqvist, K. (2013). Post-saccadic oscillations in eye movement data recorded with pupil-based eye trackers reflect motion of the pupil inside the iris. *Vision Research*, 92:59–66.
- Oakes, L. M. (2012). Advances in eye tracking in infancy research. *Infancy*, 17(1):1–8.
- Saez de Urabain, I. R., Johnson, M. H., & Smith, T. J. (2015). GraFIX: A semiautomatic approach for parsing low- and high-quality eye-tracking data. *Behavior Research Methods*, 47(1):53–72.
- Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. *Proceedings of the Eye Tracking Research and Applications Symposium*.
- Sauter, D., Martin, B. J., Di Renzo, N., & Vomscheid, C. (1991). Analysis of eye tracking movements using innovations generated by a Kalman filter. *Medical*

- Biological Engineering & Computing*, (29):63–69.
- Shic, F., Chawarska, K., & Scassellati, B. (2008). The amorphous fixation measure revisited: With applications to autism. *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*.
- Smeets, J. B. J., & Hooge, I. T. C. (2003). Nature of variability in saccades. *Journal of Neurophysiology*, 90:12–20.
- Steffen, M. (1990). A simple method for monotonic interpolation in one dimension. *Astronomy and astrophysics*, (239):443–450.
- van der Lans, R., Wedel, M., & Pieters, R. (2011). Defining eye-fixation sequences across individuals and tasks: the Binocular-Individual Threshold (BIT) algorithm. *Behavior Research Methods*, 43(1):239–257.
- Veneri, G., Piu, P., Rosini, F., Federighi, P., Federico, A., & Rufa, A. (2011). Automatic eye fixations identification based on analysis of variance and covariance. *Pattern Recognition Letters*, 32:1588–1593.
- Wang, D., Mulvey, F. B., Pelz, J. B., & Holmqvist, K. (2016). A study of artificial eyes for the measurement of precision in eye-trackers. *Behavior Research Methods*.
- Wass, S. V., Forssman, L., & Leppänen, J. (2014). Robustness and precision: How data quality may influence key dependent variables in infant eye-tracker analyses. *Infancy*, 19(5):427–460.
- Wass, S. V., Smith, T. J., & Johnson, M. H. (2013). Parsing eye-tracking data of variable quality to provide accurate fixation duration estimates in infants and adults. *Behavior Research Methods*, 45(1), 229–250.
- Zemblys, R., & Holmqvist, K. (2016). Optimal settings for commercial event detection algorithms based on the level of noise.

## 4.9. Appendix A

### 4.9.1. Noise

Here we describe how the characteristics of noise in the infant eye-movement data were determined and recreated. The following steps were performed:

1. We performed event detection on the eye movement data using I2MC (chosen for its event detection over a wide range of noise levels and data loss).
2. For each trial we selected all fixations that were at least 400 ms long.
3. For each of these fixations, we determined the frequency characteristics of the noise separately for the horizontal and vertical eye position signals as follows:
  - a) We centered and Fourier transformed the central 400 ms of the fixation using the FFT algorithm in MATLAB R2015b, yielding the frequency domain Fourier coefficients  $C_k$ , where  $k$  ranges from 0 to 60 for our 400 ms of data at 300 Hz.
  - b) We then calculated the Power Spectral Density (PSD) using standard techniques. Specifically, using the symmetry of the Fourier coefficients around the Nyquist frequency that is obtained when transforming a real signal, the Power Spectral Density (PSD) was obtained as  $C_k C_k^* / 120 \times 2$ .
  - c) Noise can be characterized by the examining how the PSD changes as a function of frequency. In biological systems, the PSD is often inversely proportional to the frequency, and as such characterized by  $1/f^a$ , where  $f$  is frequency and  $a$  a scaling exponent. Such scaling relationships have also been observed in human eye movements, with scaling exponents between 0.6 and 2 (Coey, Wallot, Richardson, & van Orden, 2012; Findlay, 1971; Wang, Mulvey, Pelz, & Holmqvist, 2016).  $a$  was determined by the

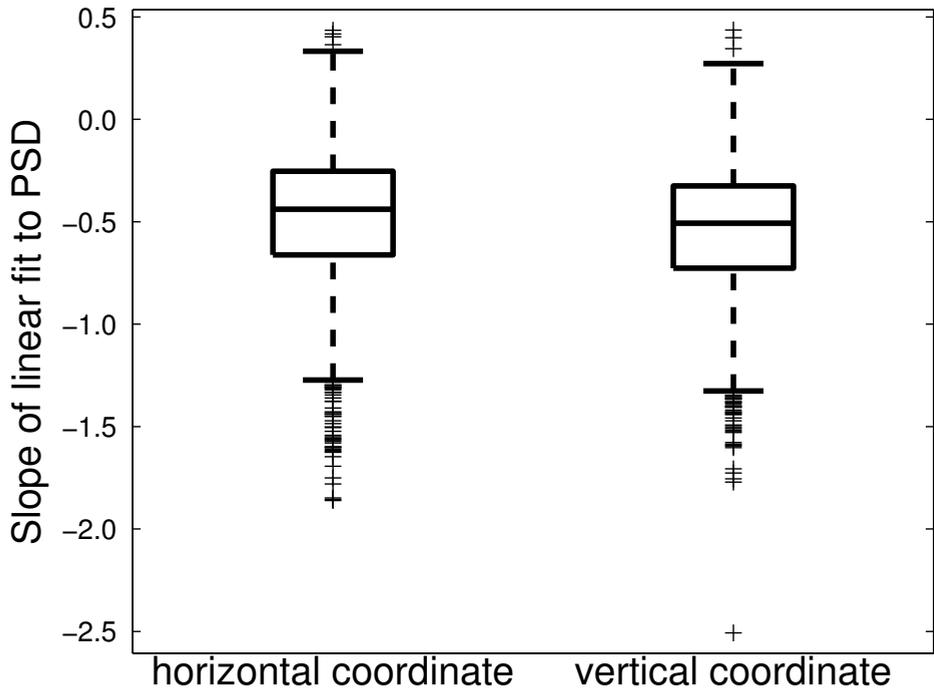


Figure 4.11.: Boxplots for the slopes of the linear fit to the power spectrum density (PSD) for the horizontal and vertical gaze coordinate.

slope of a straight line fitted to the PSD in log-log space (where a negative slope corresponds to a positive  $a$ ).

- d) The mean slope of all fixations was -0.48 for the horizontal axis and -0.54 for the vertical axis, comparable to previous results (Coey et al., 2012). However, as visible from Figure 4.11 there was considerable variation in the observed slopes. To ensure we adequately represent the observed range of slopes in the noise, we cast the fixations into 100 bins (each containing an equal number of data points) based on the slope obtained from the PSD fit. We then determined the mean PSD in each bin by taking the PSDs for all trials in the bin and averaging these.

#### 4. *Noise-robust fixation detection*

The thus derived distribution of PSDs was used to generate random noise with a sample-to-sample RMS level of between  $0^\circ$  and  $5.57^\circ$ . For each noise level, a noise signal was generated and added to each trial of the data set with the following procedure:

1. Randomly choose one of the 100 PSDs generated from the infant data set.
2. Interpolate the PSD so that it corresponds to the number of samples for which we want to generate data.
3. Generate a random signal with the same number of samples as the trial and take its Fourier transform. Combine the phase spectrum of this signal with the interpolated PSD generated in step 2.
4. Compute the random noise time series with the target PSD by taking the inverse Fourier transform of the combined signal from step 3.
5. Determine the sample-to-sample RMS of this noise time series and scale the time series to achieve the desired RMS noise level.

#### **4.9.2. Data loss**

We have previously described the empirical distribution of duration of data loss for the baby data set (Hessels et al., 2015a). To be able to generate representative data loss, we expand on this by also describing the distribution of duration of data between periods of data loss and the relationship between duration of a period of the data loss and the duration of the following period of data loss. To determine these relationships, we analyzed data loss in the baby data with the following procedure:

1. For each period of data loss, we computed its duration, the duration of data until the next period of data loss and the duration of the next period of data loss.
2. We summarized this data as a 3D histogram indicating the frequency of co-occurrence of every possible combination of the three variables.

As none of the algorithms deal with loss period longer than 100 ms, the histogram was truncated to this value for loss duration and next loss duration.

It should be noted that the above procedure describes what the loss looked like when there is loss, it does not capture how frequently loss occurs overall in the data. For instance, trials in which no loss occurs, as well as data duration before the first loss period and after the last loss period in a trial are not represented in our histogram. However, the aim is that data loss that is present is representative of how data loss occurs in sub-optimal recordings. Given the above histogram as input we sampled data loss and added it to between 0% and 100% of the trial with the following procedure:

1. For the first data loss period, randomly choose a point in the 3D histogram, taking the relative frequency of occurrence as a probability weighting for each point being selected.
2. Then, for all the next samples, use the duration of the next period of data loss as was just given by the previous sample to determine duration of the current loss period. Sample from the subset of the 3D histogram given by the current data loss duration to determine the duration of data until the next period of data loss and the duration of the next period of data loss, again using the relative frequency of occurrence of each combination as probability weights.
3. Continue with step 2 until data loss has been generated for the whole trial. Step 1 is repeated in the very rare case that there are no observations of a given loss duration in the 3D histogram. To add data to a specific percentage of the trial, we define 24 equally spaced time points in the trial. Data loss is added in windows centered on these time points, with the window size set such that the desired amount of loss is attained.

## 4.10. Appendix B

### 4.10.1. Additional algorithms for comparison

**Adaptive velocity algorithm (NH).** A recent advancement in event detection was introduced by Nyström & Holmqvist (2010). Their search rule first labels periods of data as noise when the velocity exceeds a physiological implausible value. Subsequently, a velocity threshold adapts to the remaining noise level for finding saccade-candidates, which are labeled as saccades if they exceed the minimum saccade duration (i.e. the categorization rule for saccades). Hereafter, periods of data following a saccade are labeled as post-saccade oscillations if they contain a peak exceeding a certain velocity threshold. Finally, remaining samples are labeled as fixation-candidates. If consecutive samples exceed the minimum fixation duration (i.e. the categorization rule for fixations) of 40 ms, they are labeled as a fixation. The reason for its inclusion in the present comparison is that, in our experience, it is one of the best event-detection algorithms currently available for low-noise data. We reason that, as it employs an adaptive velocity threshold set depending on the noise level in the data, the outcome measures derived from its output might be robust to noise for at least a range of noise levels.

#### **Identification by velocity threshold for low- and high-noise data (WSJ).**

Wass et al. (2013) adapted a velocity threshold search rule with several post-hoc categorization rules as detailed in the introduction. In short, the algorithm uses a fixed velocity threshold to determine fixation-candidates, and hereafter excludes those fixation-candidates that are unlikely to be genuine fixations according to a set of rules. The reason the algorithm is included here is because it was specifically designed to accomplish fixation labeling in low- and high-noise data.

**Identification by Kalman filter (I-KF).** In the I-KF algorithm a Kalman filter is used to predict eye velocity of the current sample based on the observed eye velocities of the previous samples in a noise-suppressing manner. In its search rule, the samples for which the observed velocity differs sig-

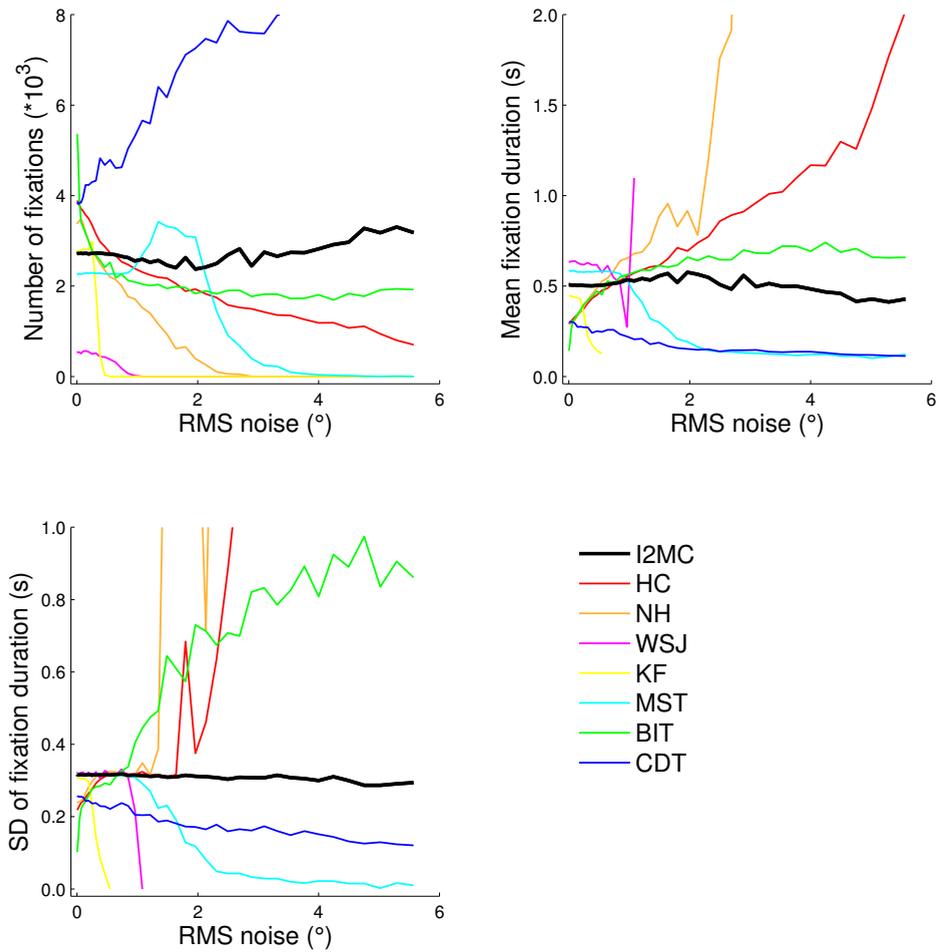


Figure 4.12.: Number of fixations (top left panel), mean fixation duration (top right panel), and standard deviation of fixation duration (bottom left panel) for all eight event-detection algorithms as a function of RMS noise added to the eye-movement data.

#### 4. Noise-robust fixation detection

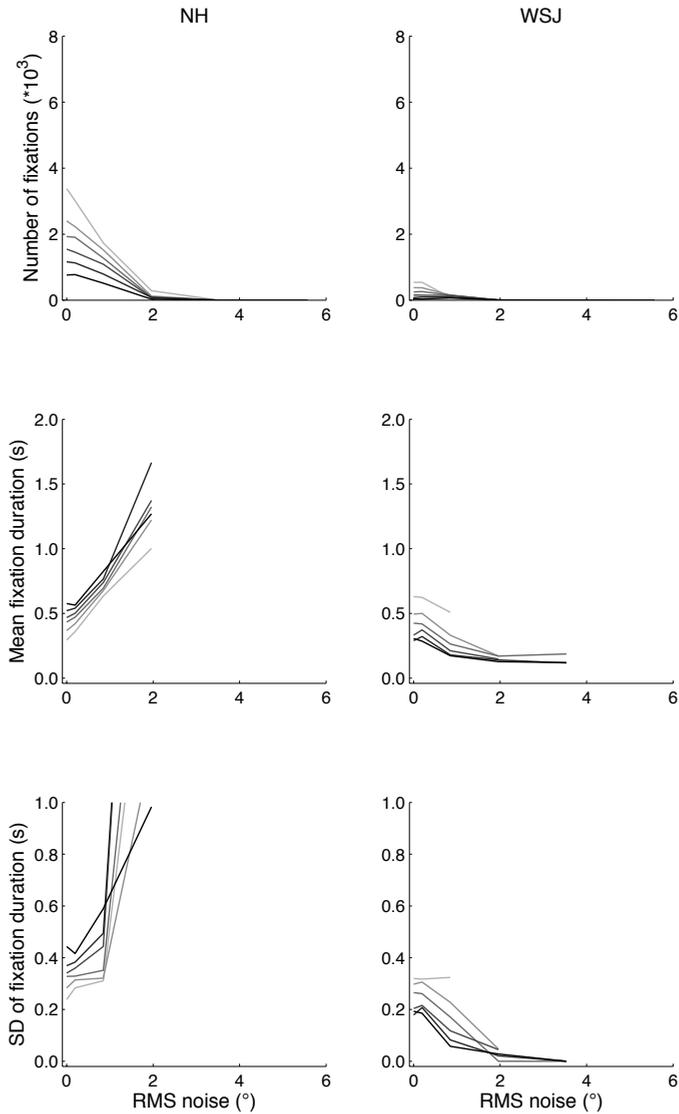
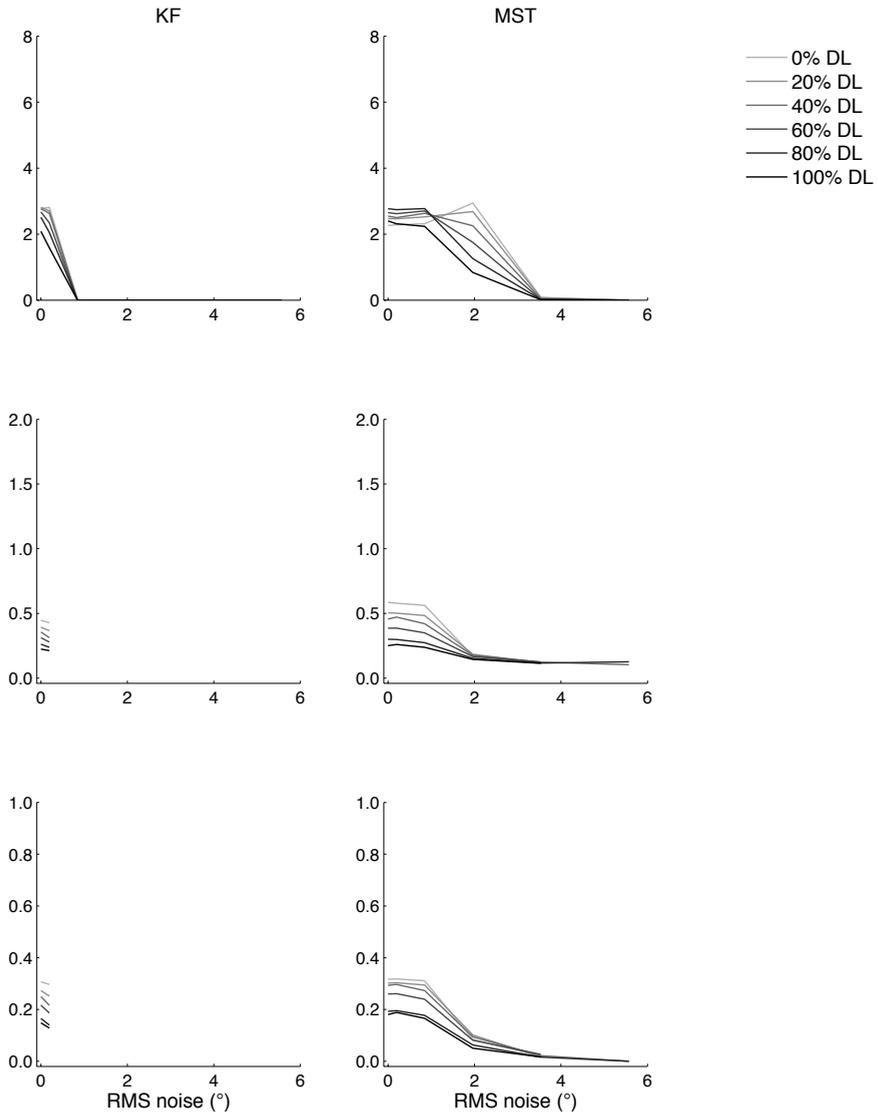


Figure 4.13.: Number of fixations (top panels), mean fixation duration (middle panels), and standard deviation of fixation duration (bottom panels) for the four additional algorithms as a function of noise level in the eye-movement data. From left to right columns depict the NH, WSJ, KF, and MST algorithms. Separate lines indicate data loss added to 0% (lightest grey) to 100% (black) of trial.

4.10. Appendix B



#### 4. Noise-robust fixation detection

nificantly from the predicted velocity (using a chi-square test) are flagged as saccade-candidates (Sauter, Martin, Di Renzo, & Vomscheid, 1991; see also Komogortsev & Khan, 2009). The implementation by Komogortsev et al. (2010) is used in the present comparison. Their categorization rules are subsequently: if fixation-candidates were separated by a Euclidean distance of less than  $0.5^\circ$  and less than 75 ms in time, fixation-candidates were merged. Fixation-candidates shorter than 100 ms were excluded.

**Identification by minimum spanning tree (I-MST).** A minimum spanning tree algorithm aims to connect all 2-d gaze coordinates with line segments in a tree in such a way that the total length of these line segments is minimized (Salvucci & Goldberg, 2000). In an ideal situation, short line segments connect gaze coordinates during a fixation, and longer line segments connect the separate fixations with each other. The reason I-MST is included here is because *“The advantage of using an I-MST is the algorithm’s ability to correctly identify fixation points even when a large part of the signal is missing due to noise”* (Komogortsev et al., 2010, p. 2638). The implementation of I-MST used in the present comparison is provided by Komogortsev et al. (2010). Their subsequent categorization rules are: if fixation-candidates were separated by a Euclidean distance of less than  $0.5^\circ$  and less than 75 ms in time, fixation-candidates were merged. Fixation-candidates shorter than 100 ms were excluded.

#### 4.10.2. Results and conclusions for additional algorithms

As can be seen in Figure 4.12, the outcome measures of the NH, WSJ, KF, and MST algorithms were not robust to increases in noise level in the eye-movement data. As visible from Figure 4.13, this was also the case for combinations of increases in noise and data loss levels. Importantly, the NH, WSJ, and KF algorithms reported little to no fixations when the noise level is over  $2^\circ$  of RMS noise. Moreover, the number of fixations, and consequently the mean fixation duration, were not robust to changes in noise level between 0 and  $2^\circ$  of RMS noise. Concluding, the four additional algorithms did not improve over the I2MC and algorithms already introduced

in the main body of the paper. Finally, Table 4.2 depicts the percentual changes in the number of fixations, mean fixation duration, and standard deviation of fixation duration for the noise, variable noise, and data loss analyses for all eight event-detection algorithms. The percentual changes in the outcome measures of the I2MC algorithm were smallest, and the I2MC algorithm is therefore considered most robust to changes in eye-movement data quality.

4. Noise-robust fixation detection

Table 4.2.: Percentage change in number of fixations, mean fixation duration, and SD of fixation duration after adding noise, variable noise, and data loss for all algorithms. \* indicates no fixations were detected.

Algorithm	Number of fixations				Fixation duration				<i>sd</i> of fixation duration			
	Noise	Var noise	Data loss	Noise	Var noise	Data loss	Noise	Var noise	Data loss	Noise	Var noise	Data loss
I2MC	17	-7	6	-16	7	-8	-7	-1	-2			
HC	-82	-77	-8	594	447	-35	1881	1598	-30			
NH	-100	-96	-76	*	-26	91	*	-38	75			
WJSJ	-100	-87	-94	*	-26	-50	*	-2	-49			
KF	-100	-50	25	*	0	-48	*	0	-48			
MST	-99	-50	7	-79	-2	-56	-97	0	-42			
BIT	-64	-84	-17	360	382	-12	743	199	-29			
CDT	144	73	62	-62	-43	-48	-53	0	-29			

## **5. The area-of-interest problem in eye-tracking research: a noise-robust solution for face and sparse stimuli**

Published as:

Hessels, R. S., Kemner, C., van den Boomen, C., & Hooge, I. T. C. (2016). The area-of-interest problem in eyetracking research: A noise-robust solution for face and sparse stimuli. *Behavior Research Methods*, 48(4):1694–1712.

Author contributions:

RH, CK, IH designed the study. CB collected the data. RH, IH analyzed the data. RH, IH interpreted the data. RH drafted the paper. RH, CK, CB, IH finalized the paper.

## **Abstract**

A problem in eye-tracking research is choosing areas of interest (AOIs): Researchers in the same field often use widely varying AOIs for similar stimuli, making cross-study comparisons difficult or even impossible. Subjective choices while choosing AOIs cause differences in AOI shape, size, and location. On the other hand, not many guidelines for constructing AOIs, or comparisons between AOI-production methods, are available. In the present study, we addressed this gap by comparing AOI-production methods in face stimuli, using data collected with infants and adults (with autism spectrum disorder (ASD) and matched controls). Specifically, we report that the attention-attracting and attention-maintaining capacities of AOIs differ between AOI-production methods, and that this matters for statistical comparisons in one of three groups investigated (the ASD group). In addition, we investigated the relation between AOI size and an AOI's attention-attracting and attention-maintaining capacities, as well as the consequences for statistical analyses, and report that adopting large AOIs solves the problem of statistical differences between the AOI methods. Finally, we tested AOI-production methods for their robustness to noise, and report that large AOIs – using the Voronoi tessellation method or the limited-radius Voronoi tessellation method with large radii – are most robust to noise. We conclude that large AOIs are a noise-robust solution in face stimuli and, when implemented using the Voronoi method, are the most objective of the researcher-defined AOIs. Adopting Voronoi AOIs in face-scanning research should allow better between-group and cross-study comparisons.

Across all fields of research using eye tracking as a research method, Areas of Interest (AOIs) are used to link eye-movement measures to parts of the stimulus used (e.g. the time spent looking at a particular object in the stimulus). AOI statistics – for example dwell time (Holmqvist et al., 2011, p. 386) – can make eye-movement data easier to interpret and are used in multiple fields of research such as user-interaction, marketing research and psychology. While it is common to provide motivation for which AOIs to construct, it is uncommon to motivate the shape and size of these AOIs. The problem that arises is that AOI statistics from different studies using similar stimuli are difficult to compare. In the present study we explore the methods and guidelines available for AOI construction, and subsequently evaluate AOI methods for use in face stimuli: A research field where researchers do not necessarily apply the same AOIs while the stimuli used are highly similar. This is particularly relevant for researchers investigating for instance (the development of) face processing, atypical face processing such as in Autism Spectrum Disorder (ASD), and social interaction using eye tracking.

Constructing AOIs, and more specifically choosing the size and shape of these AOIs, can be a difficult choice (see e.g. Goldberg & Helfman, 2010). Holmqvist et al. (2011, p. 218-219) describe that when the semantic parts of the stimulus image are clearly discernable, for instance an airplane's cockpit with separate panels and dials for separate functions, AOIs are often defined by an expert. An expert may be the researcher involved or an external expert, for instance a pilot in the previous example of an airplane cockpit. However, even if AOIs are expert-defined and locations for candidate AOIs are clear, the size and shape of a specific AOI set depend on the specific expert involved in defining the AOIs. One might wonder whether experts in the same field define the same AOIs for an identical or similar stimulus set. Moreover, constructing a sensible AOI set doesn't only warrant knowledge of the stimulus semantics (i.e. which parts of a stimulus belong to which AOI), but also of the quality of the data collected. Data quality, and more specifically, the spatial accuracy of an eye-tracking study, determines the minimum size AOIs needed in order to capture gaze

## 5. Noise-robust AOIs for face stimuli

towards that location. Data quality depends on factors such as the eye-tracking system and participant group involved in the study, and can affect AOI-based eye-tracking measures (Holmqvist, Nyström, & Mulvey, 2012). Finally, AOI production can be a laborious process when done so by hand. While there are machine-made alternatives, they are not widely applied (see Table 5.1). We will explore the methods of AOI construction used in one particular field of study: face scanning research, where experts in the same field do not define the same AOIs.

In general, AOIs are constructed separately for each study by the researcher determining the most relevant areas of the stimuli. It is therefore not surprising that studies using similar stimuli rarely have identical AOIs. One particularly striking example is in face-scanning research: While all faces have the same main features (i.e. two eyes, a nose, and a mouth), AOIs in face scanning studies vary considerably. Table 5.1 provides an overview of all the studies investigating gaze behavior with face stimuli that have provided visual examples of the AOIs used. While the eyes, nose and mouth features are always used as AOIs, rarely are AOIs identical except between studies by the same research group. In addition, the AOIs are almost always constructed using the hand-drawn method; only two studies employ (partly) machine-made alternatives. The question that arises is whether the AOI-production method or AOI set used matters for the AOI-based measures used in analyses. If it doesn't, one might consider applying the method with the least amount of labor involved. If it does, the question arises which production method or AOI set is preferable in a specific situation or for a specific AOI-based measure.

Table 5.1.: AOI methods used in face scanning studies. Only studies reporting visual examples of the AOIs used are included. The methods of AOI construction are elaborated on in Table 5.2. Note that both the Voronoi-tessellation and Limited-radius Voronoi-tessellation AOI methods are absent in the literature. We are not aware of any face scanning studies applying these methods.

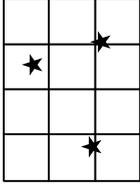
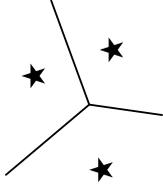
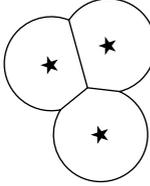
Study	Method	AOIs of inner facial features	Fixed AOI shape	Whitespace between AOIs
Hunnius & Geuze (2004)	Grid	Eyes and mouth AOIs constructed from multiple grid cells.	No: AOIs are created from several grid cells	No
Gallay et al. (2006)	Grid & Hand-drawn	Grid: no inner feature AOIs constructed from grid cells. Hand-drawn: Eyes and nose/mouth area constructed.	Yes: rectangles	No
Nguyen et al. (2009)	Hand-drawn	Eyes, upper lip, lower lip, nose, glabella, forehead, cheek	No	Yes
Senju et al. (2013)	Hand-drawn	Left eye, right eye, nose bridge, mouth, nose	Yes: rectangles	No
Oakes & Ellis (2013)	Hand-drawn	Eyes, mouth, nose	Yes: rectangles	No
Tenenbaum et al. (2013)	Hand-drawn	Eyes, mouth, nose	Yes: rectangles	No
Võ et al. (2012)	Hand-drawn	Eyes, mouth, nose	Yes: rectangle (mouth) and ellipses (other AOIs)	Yes
Kano & Tomonaga (2010)	Hand-drawn	Eyes, mouth, nose	No	No
Chawarska & Shic (2009)	Hand-drawn	Eyes, mouth, nose	No	No
Shic et al. (2014)	Hand-drawn	Eyes, mouth, nose	No	No
Liu et al. (2011)	Hand-drawn	Eyes, mouth, nose	No	Yes
A. Wheeler et al. (2011)	Hand-drawn	Eyes, mouth, nose	No	Yes

## 5. Noise-robust AOIs for face stimuli

Falck-Ytter (2008)	Hand-drawn	Eyes, mouth	Yes: rectangles and ellipses	Yes
Wagner et al. (2013)	Hand-drawn	Eyes, mouth	Yes: rectangles	Yes
Rutherford & Towns (2008)	Hand-drawn	Eyes, mouth	Yes: rectangles	Yes
Wilcox et al. (2013)	Hand-drawn	Eyes, mouth	Yes: rectangles	Yes
Jones et al. (2008)	Hand-drawn	Eyes, mouth	No	Yes
Jones & Klin (2013)	Hand-drawn	Eyes, mouth	No	Yes

There are several methods available for AOI construction; we limit the scope to researcher-defined AOIs, i.e. the expert-defined AOI category as described by Holmqvist et al. (2011). While there are multiple other methods available, including data-driven alternatives to AOI analyses (see e.g. Caldara & Mielle, 2011), we are concerned here with defining AOIs when the features to cover by the AOIs and the hypotheses concerning them are clear. We address additional approaches in more detail in the discussion. When considering researcher-defined AOIs, there are multiple methods for AOI construction that differ in a number of ways. AOIs can be implemented fully by the researcher (which we will refer to as man-made) or they can be partly implemented by custom software (machine-made). Moreover, AOIs can differ in whether they are subjective in location, shape or size (see Table 5.2). One approach often used is the manual selection of an area around a part of the stimulus that is of interest to the researcher, for instance a hand-drawn ellipse around the eye in a static face. While the specific implementation of creating such AOIs might be limited to certain shapes (e.g. software for constructing AOIs that allows only rectangles and ellipses versus all forms), this method is most akin to drawing areas on a printout of a stimulus and will therefore be referred to as the hand-drawn AOI method. Using the hand-drawn method, the location, size and shape of an AOI are subjective choices, and the AOIs are man-made. A different method that has previously been used in the face scanning literature is the grid method. Using this method, a grid is placed over the stimulus and each grid cell is considered an AOI. Subsequently grid-cells can be attributed to features (e.g. column 2, row 1 is part of the ‘eyes’ AOI), but this is not a requirement. Each grid-cell AOI has an objective location and shape, yet AOI size is subjective. The grid method is less laborious to implement than the hand-drawn method; it can be done partly by machine by computing the locations of grid cells given a cell size. However, if one is interested in calculating feature-specific measures (i.e. total time spent looking at the eyes of a face), the grid method still needs grid-cells to be attributed to separate AOIs. If grid-cells are attributed to specific AOIs, the location and shape of that specific AOI become a subjective choice.

5. Noise-robust AOIs for face stimuli

AOI method	Location	Shape	Size	Man-/machine-made	Example
Hand-drawn	Subjective	Subjective <sup>1</sup>	Subjective	Man-made	
Grid	Objective / Subjective <sup>2</sup>	Objective / Subjective <sup>2</sup>	Subjective	Machine-made	
Voronoi-tessellation	Subjective	Objective	Objective	Machine-made	
Limited-radius Voronoi-tessellation (LRVT)	Subjective	Objective <sup>3</sup>	Subjective	Machine-made	

- 
- <sup>1</sup>Depending on the implementation of AOI construction restrictions on shape may occur, e.g. custom scripts allowing all shapes versus eye-tracker manufacturer software allowing only certain shapes (e.g. rectangles or ellipsoids). These restrictions are at the level of implementation and not at the conceptual level and will therefore be ignored.
- <sup>2</sup>If the grid cells are analyzed as is, and not assigned to specific AOIs location and shape are objective. If grid cells are, however, assigned to specific AOIs, a subjective choice has to be made, which makes the location and shape constructed from the grid cells subjective.
- <sup>3</sup>While one might subjectively choose to use the LRV method and thus the shape of a circle – particularly when there are no Voronoi borders within the chosen radius – the circle-shape of the AOI is not based on the stimulus on which the AOI is constructed and therefore objective with regard to stimulus content.

## 5. Noise-robust AOIs for face stimuli

We present here two related AOI methods based on the tessellation principle introduced by Voronoi (1909), both of which are currently unused in face scanning research. The first method, the Voronoi-tessellation method (or Voronoi method for brevity), can be used to divide an area around a number of points. Each cell defined by the Voronoi method represents the area that is closest to one of the points (i.e. its cell center), and lines indicate locations where any of two points, or cell centers, is at equal distance. Each fixation that falls within a Voronoi AOI describes that fixation as being closer to its cell center, for example the center of the nose, than any of the other cell centers (the centers of one of the eyes or the mouth). Figure 5.1 demonstrates an example of Voronoi tessellation with three cell centers. In eye-tracking, the Voronoi method has previously been used to quantify distributions of fixations (Over, Hooge, & Erkelens, 2006), as AOIs for calibration/validation targets (Nyström, Andersson, Holmqvist, & van de Weijer, 2013), and they have been applied to many different problems (see Aurenhammer, 1991, for an overview). The Voronoi method includes all space of the stimulus, which in turn means that there is no area of the stimulus left that does not belong to an AOI. After selection of cell centers the construction of AOIs can be done by machine and both shape and size are objective. A variation of the Voronoi method is what we will call the Limited-radius Voronoi-tessellation method (LRVT method for brevity), the last method we will review. Instead of dividing the stimulus-space based on the cell centers, the LRVT method uses cell centers and a given radius to produce AOIs. Each fixation that falls within a LRVT AOI describes that fixation as being both closest to that AOI's cell center and being within a given radius from that cell center. This method could be used if one would want to use a similar method as the Voronoi method, yet not include all whitespace (i.e. space not part of the stimulus, but included on the screen) around the stimulus. Not including all whitespace might be useful if there is a lot of gaze outside the stimulus of interest, for instance directed at a cursor remaining on screen, or an object close to the border of the screen. Like the Voronoi method, the LRVT method can be implemented using a machine, yet both location and radius (and thus AOI-size)

are subjective choices. We are not aware of any studies on face scanning using the Voronoi or LRVT method.

Although no study has directly compared different AOI methods, there are several researchers that have suggested guidelines for AOI construction. For instance, Goldberg & Helfman (2010) suggest that AOIs should only be defined for objects of interest: they need not fill the entire screen. They furthermore suggest that the padding (or margin) around an object should depend on three factors: “(1) the importance of capturing every fixation on that object, (2) the amount of white space surrounding the object, and (3) expected variance in fixation positions across participants” (p. 72). Holmqvist et al. (2011) suggest that, when possible, objects of interest should be positioned in the stimulus such that white space between AOIs can exist (i.e. AOIs should not be contiguous). In addition, they outline that the minimal AOI size should be determined by the precision and accuracy of the recorded eye-tracking data. Finally, Holmqvist et al. (2011) suggest that arbitrary AOI positioning should be avoided: they should be as precise as possible with regard to the objects of interest in the stimulus. Hooge & Camps (2013) add that for sparse stimuli – relatively empty stimuli where there is not much crowding (lateral masking) such as faces – AOIs should be as large as possible. Objects of interest are visible from a larger eccentricity in sparse stimuli than in dense stimuli. Concluding, AOI construction should depend on both data quality and the type of stimulus used.

In the present study we explore how different AOI methods affect eye-tracking measures. We evaluate the AOI methods on measures for two AOI characteristics that are commonly used in eye-tracking research: attention-attracting, and attention-maintaining capacity. These two AOI characteristics are being evaluated for the following reasons: We first assume that participants will direct their gaze to AOIs with a high attention-attracting capacity sooner in time than to AOIs with a lower attention-attracting capacity. Attention-attracting capacity might be useful for marketing researchers investigating the time it takes customers to look at a brand label,

5. Noise-robust AOIs for face stimuli

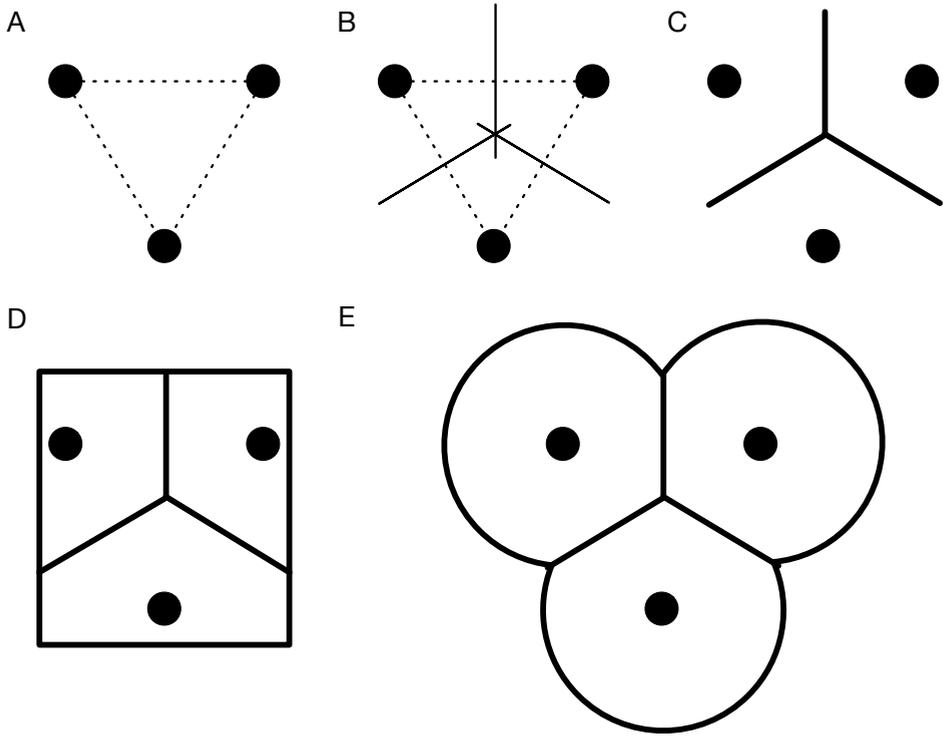


Figure 5.1.: Explanation of the Voronoi tessellation method. (a) Determine the cell centers and draw connecting lines between them. (b) Draw perpendicular bisection lines for each connecting line. (c) Find borders between the cells by connecting the bisection lines. Following Voronoi tessellation, AOIs may be confined: for instance, within the screen or stimulus dimensions (d) or by allowing a maximum radius, as in the limited-radius Voronoi tessellation method (e). For an elaborate example of Voronoi tessellation using more cells, see Over, Hooge, & Erkelens (2006)

or for experimental psychologists investigating reaction times to peripheral targets. The second assumption is that participants will retain their gaze for a longer time in AOIs with high attention-maintaining capacity compared to AOIs with a lower attention-maintaining capacity. Attention-maintaining capacity might, for instance, be useful for developmental psychologists investigating infant preferences for objects or faces.

We focus specifically on three questions regarding attention-attracting and attention-maintaining capacities of AOIs. 1) How do the attention-attracting and attention-maintaining capacities of AOIs differ between AOI-production methods? 2) What is the relation between AOI size and AOI attention-attracting and attention-maintaining capacity? Are there possible implications thereof for statistical comparisons within and between groups? 3) Which AOI type is most robust to noise? If attention-attracting and attention-maintaining capacities differ between AOI-production methods and are dependent on AOI size, care should be taken when comparing results across studies using different AOIs. In addition, robustness to noise is important to consider when different participant groups are compared in the same study, or when two studies using eye-trackers with different noise levels are compared. In infant eye-tracking research, for example, data are typically noisier compared to adult eye-tracking research. We therefore investigate these three questions in data collected in three different participant groups: typically developing adults, adult with autism spectrum disorder (ASD), and infants. All datasets were obtained using face stimuli, and we will therefore limit ourselves to AOI methods for face scanning studies. However, as faces are sparse stimuli, our observations might very well generalize to a broader range of studies using sparse stimuli – for example studies investigating gaze behavior to advertisements or psychological displays, both of which are often sparse. The ASD and typically developing adults, as well as the infant participant group were chosen as they are of particular relevance in the face scanning literature. See Guillon, Hadjikhani, Baduel, & Rogé (2014) for a recent review on face scanning in ASD, and e.g. Hunnius, de Wit, Vrins, & Hofsten (2011); Wilcox, Stubbs, Wheeler, & Alexander (2013) for face scanning in infancy.

## 5.1. Methods

### 5.1.1. Participants

#### Dataset 1 - infants

40 infants (19 male, 21 female) participated in the present study. Mean age of the included group was 311 days ( $sd = 16.7$  days). All were born full-term (38-42 weeks), had normal birth weight, and no delays in development or abnormalities in visual or auditory processing were reported by the health-care system. The medical ethical committee of the University Medical Center Utrecht approved the study. All parents or caretakers gave written informed consent prior to participation, after explanation of the procedure.

#### Dataset 2 - ASD and matched controls

13 young adults with ASD (11 male, 2 female) and 16 matched control subjects (13 male, 3 female) participated in the experiment. 3 subjects with ASD (2 male, 1 female) were excluded from the analysis for either skipping through the trials in more than 20% of the trials (i.e. making only one fixation,  $n = 2$ ) or technical difficulties with the eye-tracker ( $n = 1$ ). 5 control subjects (3 male, 2 female) were excluded from the analysis for either skipping through the trials ( $n = 3$ ), or not fixating in the middle of the screen between trials more than 20% of the experiment ( $n = 2$ ). Descriptive statistics of the included group are given in Table 5.3. For the ASD group, the Wechsler Adult Intelligence Scale III, Dutch edition (WAIS-III), was used to determine IQ scores. For the control group the Wechsler Abbreviated Scale of Intelligence (WASI) was used to estimate IQ. The diagnostic evaluation for the ASD group included a psychiatric observation and a review of prior records (developmental history, child psychiatric and psychological observations and tests). ASD was diagnosed by a child psychiatrist using the DSM-IV criteria. The medical ethical committee of the University Medical Center Utrecht approved the study.

Table 5.3.: Descriptive statistics for the ASD and control group. Standard deviation is given in parentheses.

	ASD group	Control group
Sample size	10	11
Age	23.2 (4.25)	23.3 (2.33)
Full-scale IQ	113.3 (9.36)	120.3 (11.67)
Verbal IQ	114.6 (9.96)	121.5 (10.14)

### 5.1.2. Apparatus and stimuli

An HP EliteBook 8560w was used to present stimuli to a 23" external screen (Tobii screen attached to the eye-tracker) at a resolution of 1920 by 1080 pixels. During the task eye-movements were recorded using the Tobii TX300 at 300Hz, capable of recording at 0.4° accuracy (binocular), and 0.15° precision (unfiltered) under ideal conditions.

The stimuli consisted of 12 (6 identities) static pictures of faces taken from the MacBrain Face Stimulus Set<sup>4</sup>. The pictures were cropped and de-colored (i.e. turned to gray scale pictures). The set contained 3 female and 3 male faces, each displaying a neutral and fearful expression. The stimuli were provided by de Jong, van Engeland, & Kemner (2008). Each face measured 16.7° by 11.5° of visual angle on screen.

Each face was used in two conditions; a high contrast and a low contrast condition, by manipulating the contrast of the eye region. For the infant group, each processed face was presented once resulting in 24 trials; six neutral low contrast, six neutral high contrast, six fearful low contrast & six fearful high contrast faces. For the ASD and matched controls group, each processed face was presented eight times resulting in 192 trials; 48 neutral low contrast, 48 neutral high contrast, 48 fearful low contrast & 48

<sup>4</sup>Development of the MacBrain Face Stimulus Set was overseen by Nim Tottenham and supported by the John D. and Catherine T. MacArthur Foundation Research Network on Early Experience and Brain Development. Please contact Nim Tottenham at tott0006@tc.umn.edu for more information concerning the stimulus set.

## 5. Noise-robust AOIs for face stimuli

fearful high contrast faces. As the present study is aimed generally at AOIs in face stimuli, results from the different contrast conditions and emotional expressions are pooled.

### 5.1.3. Procedure

#### Dataset 1 - infants

The experiment took place either at the infants' home ( $n = 32$ ) or in the lab ( $n = 8$ ). If the experiment took place at home, a tent was placed over the dining table to approximate equal lighting conditions for each measurement. This tent was specifically designed for conducting research on infants at home. It included fabric in front, to the left and right sides, and above the child to block surrounding visual information. The back of the tent was left open, such that parents or experimenters could be with the child if it felt uncomfortable on its own. For all measurements, the Tobii TX300 was placed on the table, and the infant was placed in a Bumbo seat fitted with a Bumbo Playtray such that the distance between the infants' eyes and the eye-tracker was 65 cm.

The experiment was preceded by a five-point calibration sequence, after which individual points were recalibrated if necessary. The experiment began when the experimenter deemed the calibration sufficient enough. Each trial began with a colorful movie accompanied by sound, which remained on screen until the infant fixated the screen and the experimenter pressed a button to continue. After this, one of 24 face stimuli was presented for five seconds, after which another movie appeared on screen. The experiment lasted approximately 10 minutes including calibration.

#### Dataset 2 - ASD and matched controls

Participants were brought into the lab where they were positioned behind the Tobii TX300. After a built-in nine-point calibration, the experiment began. Each trial began with a black fixation cross, and participants were instructed to press the spacebar to initiate a trial. Hereafter the fixation

cross changed color and then remained on screen for a variable amount of time (1, 1.25, or 1.5 s). Subsequently, one of the 24 distinct face stimuli was presented, which participants were instructed to look at. A face was presented for a maximum duration of four seconds, or until the participant decided he/she had seen enough. The experiment for the ASD and matched controls group was self-paced, as we did not want the participants to wonder where to look when they had scanned the stimulus. Participants were not given a task (i.e. free-viewing). For both datasets the order of pictures was mixed randomly with a set of restrictions using Mix (van Casteren & Davis, 2006). Faces of at least two other identities interleaved faces of the same identity. In addition, the maximum number of repetitions of the same emotion or same contrast level was set at three.

#### 5.1.4. Data reduction

Raw position signals from the left and right eye were first combined into an average position signal. If gaze position was only available from one eye, that signal was used. Hereafter, a fixation detection algorithm specifically designed for use across varying noise levels – from low noise in adult data, to higher noise in infant data – was applied. The algorithm operates as an adaptive dispersion algorithm, with which fixation detection can be achieved across larger variations in noise levels, both within and between participants or trials. The algorithm, Identification by 2-Means Clustering (I2MC), is based on a procedure called k-means clustering (where  $k = 2$ ), which is used to determine whether one or two fixation clusters are present in a small moving window. As the I2MC algorithm employs a moving window in which clustering is carried out, it is robust to variations in local noise. In the present study we used a moving window of 200 ms width.

After event detection, RMS noise for each fixation was calculated as an estimate for precision of the recording (Holmqvist et al., 2011, p. 35). While the Tobii TX300 is capable of recording with  $0.15^\circ$  precision, this does not represent the normal value obtained in most eye-tracking studies. As can be seen from the histograms in Figure 5.2, the distribution of RMS noise in

## 5. Noise-robust AOIs for face stimuli

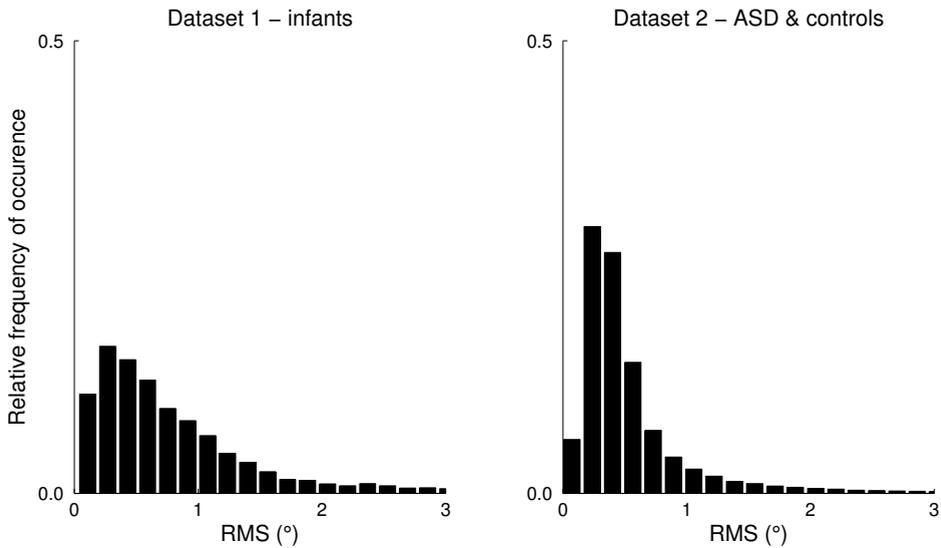


Figure 5.2.: Histograms of the root-mean square (RMS) noise, in degrees, during fixations for the infant dataset (left panel) and for the autism spectrum disorder (ASD) and matched controls dataset (right panel)

the adult dataset was narrow compared to the infant dataset. Higher RMS noise was relatively more common in the infant dataset, as visible from the slightly longer right tail of the distribution.

### AOI span

In the present study we investigate, amongst other factors, AOI size and AOI robustness to noise. While these are commonly noted in degrees of visual angle, they are difficult to relate to the stimulus used. Knowing for example that the size of an AOI is  $2^\circ$  by  $2^\circ$  is meaningless unless we specify the distance between two AOIs: the information is stimulus- and setup-specific. In order to make interpretation of the results presented here easier – especially when relating the findings to the stimuli used – we present another measure: AOI span. AOI span is the mean distance from each AOI cell center to the cell center of its closest neighbor. For example,

if all AOIs in a stimulus are circle-shaped and have a radius of 0.5 AOI, and are positioned at equal distance from each other, the borders of AOIs connect. If, as another example, the location of a fixation on the mouth were moved upward with 1 AOI span, its new location would now roughly be on the nose. For the present study, AOI span was calculated as follows. The closest AOI for both the left and right eye was the nose (at a distance of  $4.4^\circ$ ). The closest AOI for the nose was the mouth (at a distance of  $3.5^\circ$ ), and vice versa. As such, AOI span was  $3.95^\circ$ . Values of distances relating to the stimulus will henceforth be given in AOI span, with degrees given in parentheses.

### **AOI methods**

Hand-drawn AOIs were manually defined for each individual face using a standard graphics editor (Adobe Photoshop). Areas were defined for the left eye, right eye, nose, and mouth. A non AOI was included to capture all gaze data not in the feature AOIs. For the Voronoi AOI method, cell centers were defined in each individual face by determining the center of the pupils, tip of the nose, and the center of the mouth. A non AOI was included for all gaze data not on the screen (i.e. an infant looking away from the screen). If a fixation was at equal distance from 2 or more AOIs (i.e. exactly on the border), it was added to the non AOI. For the LRVT method, cell centers from the Voronoi method were used. A non AOI was included for all gaze data outside the LRVT radii, as well as for fixations at equal distance from 2 or more AOIs. We calculated eye-tracking measures for the LRVT method with a radius of 0.6 AOI span (corresponding to  $2.3^\circ$ ). While this value is partly arbitrary, the largest differences in eye-tracking measures between methods occur for radii around this value. In addition, a 0.6 radius produces AOIs that are just contiguous. We explore the effect of varying LRVT radius in depth in question 2 of the results section. The grid method was applied by centering a grid of 19 columns by 10 rows (each cell measuring 0.6 by 0.6 AOI span;  $2.3^\circ$  by  $2.3^\circ$ ) in the screen (see Figure 5.3 for an example). Grid cells were subsequently assigned to a left eye (4 cells), right eye (4 cells), nose (4 cells), mouth (6 cells), or non AOI. Cells

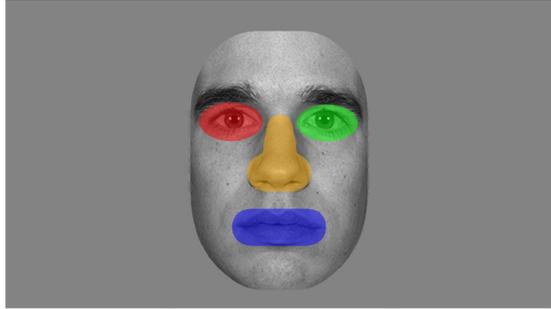
## 5. Noise-robust AOIs for face stimuli

assigned to each AOI were identical across stimuli.

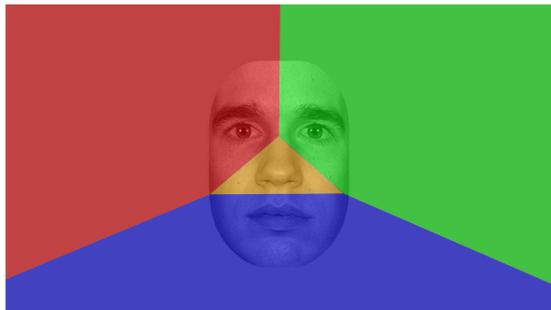
### Eye-tracking measures

As previously described, we examine the effect of using different AOI methods on two specific characteristics of an AOI: attention-attracting, and attention-maintaining capacity. To measure attention-maintaining capacity, dwell time and total dwell time were calculated. Dwell time is the time that gaze remains in a particular AOI, from entry to exit (Holmqvist et al., 2011, p. 386–389). Dwell time can only be calculated for trials in which time was actually spent in that AOI (i.e. a dwell time of 0 ms cannot occur). Mean dwell time reflects the average time that gaze remains in a particular AOI each single period. AOIs with a higher mean dwell time are assumed to maintain attention for longer individual periods than AOIs with a lower mean dwell time. Total dwell time, on the other hand, is the total time in a whole trial that gaze was in a particular AOI (Holmqvist et al., 2011, p. 389). Total dwell time can be calculated for each trial, regardless of an AOI was fixated or not. This means that total dwell times of 0 ms in a trial can also occur. AOIs with a higher mean total dwell time are assumed to maintain attention longer overall than AOIs with a lower mean total dwell time. To measure attention-attracting capacity, the time to first AOI hit was calculated; the latency from trial onset to the time where gaze first enter a particular AOI (Holmqvist et al., 2011, p. 437). Note that attention-attracting and attention-maintaining capacity are AOI-specific. Changing the size or location of an AOI results in a different AOI, and the amount of data that it includes changes also. By changing the size or location a new AOI has been created with its own attention-attracting and attention-maintaining capacity. While two AOIs from different AOI-methods may share the name (i.e. the left eye AOI for the LRVT and the hand-drawn method), they differ in the amount of data they include and thereby their attention-attracting and attention-maintaining capacities, as estimates for the underlying stimulus feature that they aim to cover.

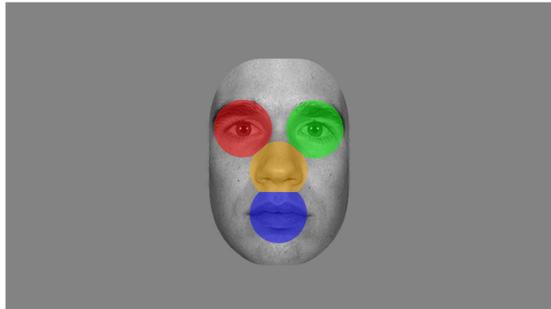
Hand-drawn



Voronoi



LRVT



Grid

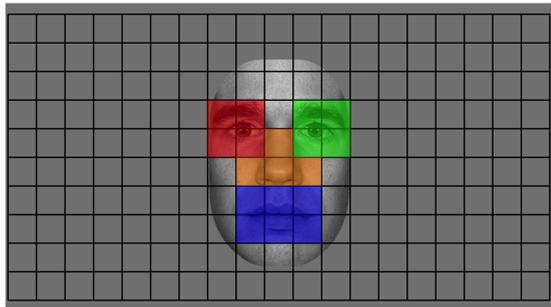


Figure 5.3: Example of the AOIs used for one stimulus in the present study. Color-coding for the left eye (red), right eye (green), nose (orange), and mouth (blue) AOIs is maintained throughout the chapter

## 5.2. Results

### 5.2.1. Question 1 – How do the attention-attracting and attention-maintaining capacities of AOIs differ between AOI-production methods?

Raw scores and group means for dwell time, total dwell time, and time to first AOI hit were plotted for each AOI-production method in the left panels of Figure 5.4 for dataset 1 (infants), and in the middle and right panels in Figure 5.4 for dataset 2 (ASD and matched controls). As can be seen from the left panels in Figure 5.4, there was some variation in group means between AOI-production methods for the infant participants. The relative pattern of means to feature AOIs (eyes, nose, and mouth) per AOI-production method was relatively consistent. Mean dwell time and mean total dwell time were consistently longer for the eye AOIs compared to the nose and mouth AOIs. Mean time to first AOI hit was consistently shortest for the nose AOI, followed by the eye AOIs and finally the mouth-AOI. (Total) dwell times to the non AOI seemed to be inversely related to feature AOI size. This was expected, as larger feature AOIs result in a smaller non AOI. For instance, they were shortest for the Voronoi method, for which only data outside the screen were labeled as belonging to the non AOI. Time to first AOI hit for the non AOI seemed related to feature AOI size: the smaller the feature AOIs, the shorter the time to first hit of the non AOI.

As can be seen from the middle and right panels in Figure 5.4, the pattern observed for group means between AOI-production methods for the ASD and matched controls was again consistent. Mean dwell times for the ASD group were consistently slightly shorter for the eye AOIs compared to the nose and mouth AOIs for all AOI-production methods. For the control group, mean dwell times for all feature AOIs were comparable. Mean total dwell times for the ASD group were consistently shorter for the eye AOIs and mouth AOI compared to the nose AOI for all AOI-production methods. For the control group, mean total dwell times are shortest for the mouth

AOI, followed by the eye AOIs and nose AOI (although there was some variation in the order of the latter two over AOI-production methods). Mean time to first AOI hit for the feature AOIs was highly consistent over AOI-production methods. For both the ASD and control group, the order from first to last was nose, left eye, right eye, and finally mouth. There were, however, absolute differences in mean to first AOI hit between groups. (Total) dwell times to the non AOI again seemed to be inversely related to feature AOI size. They were shortest for the Voronoi method, for which only data outside the screen were labeled as belonging to the non AOI. Time to first AOI hit for the non AOI seemed again related to feature AOI size: the smaller the feature AOIs, the shorter the time to first hit of the non AOI. This pattern for time to first AOI hit as a function of AOI size was, however, less consistent for dataset 2 compared to dataset 1.

The indication from these results would be that although there was some variation between group means across AOI-production methods, the relative pattern remained similar across AOI-production methods, for both the infant and for the ASD and matched controls dataset. The absolute differences appeared largely due to AOI size, which is further addressed in Question 2. To investigate whether relative differences in total dwell time to AOIs between AOI-production methods affected statistical outcome we carried out paired t-tests on mean total dwell time to both eyes (i.e. the sum of total dwell time to the left eye and right eye) and mouth. This specific comparison was carried out, as it is a recurring hypothesis with mixed results in the ASD literature (see e.g. Guillon et al., 2014). Paired t-tests for both the infant and control participants revealed that for all AOI-production methods mean total dwell time to the eyes was significantly longer than to the mouth ( $p < 0.05$ ). For the ASD group mean total dwell time to the eyes was significantly longer than to the mouth for the Hand-drawn, Voronoi and LRVT AOI-production methods ( $p < 0.05$ ), but not for the grid method ( $p > 0.10$ ). This indicates that the differences in eye-tracking outcome measures due to the AOI-production method used affected the outcome of a hypothesis-driven experiment. We examined the relationship between AOI size and attention-attracting and

## 5. Noise-robust AOIs for face stimuli

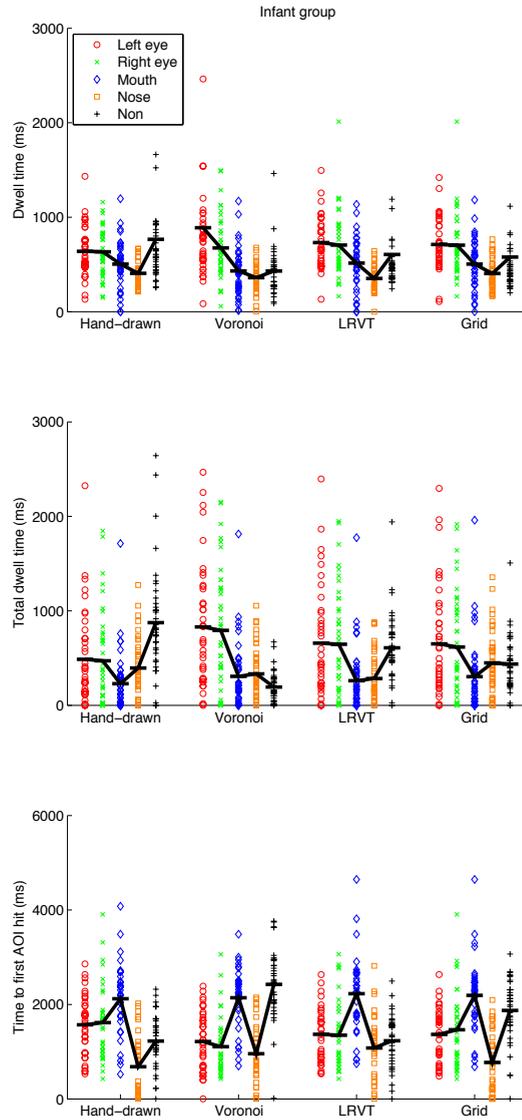
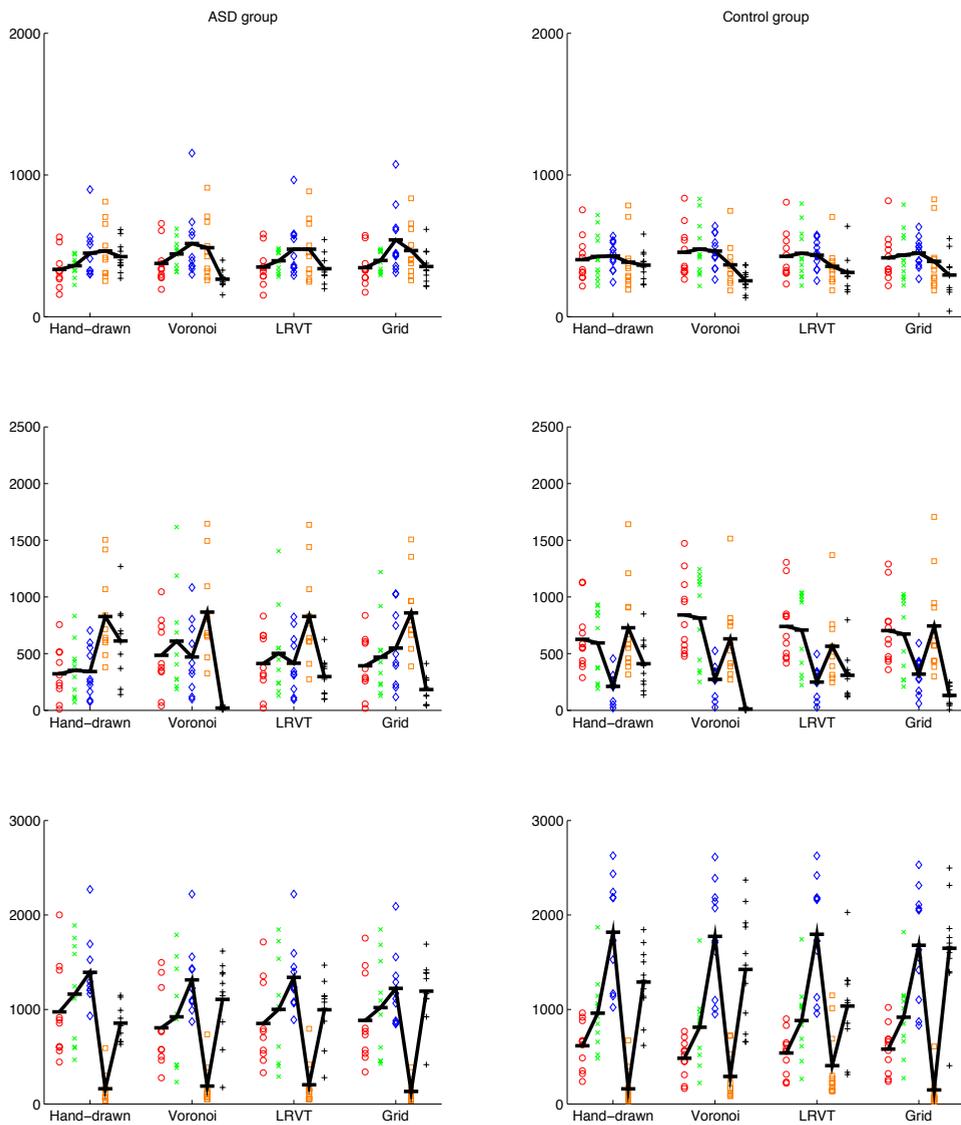


Figure 5.4.: Participant means for dwell time (top panels), total dwell time (middle panels), and time to the first AOI hit (bottom panels), separated by AOI-production methods. Black horizontal bars indicate the group means, and black connecting lines between the AOIs are added to facilitate pattern comparisons across the AOI-production methods. Panels in the left column are for the infant group, panels in the middle column for the ASD group, and panels in the right column for the control group.

## 5.2. Results



## 5. Noise-robust AOIs for face stimuli

attention-maintaining capacity, as well as the effects on statistical analyses, next.

### 5.2.2. Question 2 – What is the relation between AOI size and AOI attention-attracting and attention-maintaining capacity?

The relation between AOI size and AOI attention-attracting and attention-maintaining capacity was investigated by applying the LRVT method with a range of radii from 0.05 to 1.52 AOI span ( $0.2$ - $6.0^\circ$ ) for the feature AOIs. The LRVT method was chosen because it allows for an automated implementation compared to e.g. using increasingly larger hand-drawn AOIs. The maximum of 1.52 ( $6^\circ$ ) was chosen as this was the mean eye-to-eye difference in the stimulus, and 1.5 times the AOI span. Increasing the LRVT radius results in larger AOIs, hence LRVT radius is analogous to AOI size. For the non AOI, we generally expected the inverse relation of that observed for the feature AOIs. As size for the feature AOIs increases, size for the non AOI decreases.

Figure 5.5 depicts the relation between mean dwell time, mean total dwell time, and mean time to first AOI hit and LRVT radius for the infant dataset. As can be seen from the top left panel in Figure 5.5, mean dwell time to the eye AOIs increased with increasing LRVT radius up to around 0.75 AOI span ( $\approx 3^\circ$ ). For the mouth and nose AOIs there was no increase in mean dwell time with increasing LRVT radius. Mean dwell time for the non AOI decreased up to 0.75 AOI span ( $3^\circ$ ) LRVT radius. As can be seen from the top right panel in Figure 5.5, mean total dwell time to the feature AOIs increased with LRVT radius, again up to around 0.75 AOI span ( $3^\circ$ ). Hereafter the mean total dwell times to the feature AOIs did not increase substantially. As expected, the inverse relation was found for the non AOI: mean total dwell time decreased with increasing LRVT radius for the feature AOIs. Note that, unlike dwell times, total dwell times began at 0. This is expected, as dwell times can only be calculated for trials in which a participant fixated an AOI, whereas total dwell times are 0 when

participants did not fixate an AOI.

The mean time to first AOI hit (bottom left panel in Figure 5.5) slightly decreased with LRVT radius for the eye and nose AOIs, whereas it remained relatively stable for the mouth AOI. Mean time to first hit for the non AOI increased as LRVT radius for the feature AOIs increases. The bottom right panel in Figure 5.5 depicts the number of participants for which a time to first AOI hit could be calculated. Time to first AOI hit can only be calculated if an AOI was actually fixated in a trial. As can be seen the number of participants included in the mean time to first AOI hit for the feature AOIs increased sharply between 0 and 0.5 AOI span ( $0-2^\circ$ ) of LRVT radius.

Figure 5.6 depicts the mean dwell times, mean total dwell times, and mean time to first AOI hit for all AOIs, separated for the ASD and control group. The number of participants that were included in the calculation of time to first AOI hit is not depicted, as almost all participants were included from the smallest LRVT radius onwards. This was a result of more trials being presented compared to dataset 1.

As can be seen from the top panels in Figure 5.6, mean dwell time to all feature AOIs increased slightly with LRVT radius for both the ASD and control group, whereas mean dwell time to the non AOI decreased sharply up to 0.75 AOI span ( $3^\circ$ ) LRVT radius. As visible from the middle panels in Figure 5.6, mean total dwell time to the feature AOIs increased with LRVT radius, but approached an asymptote after 0.75 AOI span ( $3^\circ$ ) of LRVT radius. Although the mean total dwell times to feature AOIs differed between groups, the point at which the increase in total dwell time leveled off did not differ between groups. Mean total dwell time for the non AOI decreased sharply up to an LRVT radius of 0.75 AOI span ( $3^\circ$ ). Mean time to first AOI hit for all feature AOIs decreased with LRVT radius, but approached an asymptote after an LRVT radius of 0.75 AOI span ( $3^\circ$ ). Mean time to first AOI hit for the non AOI showed the inverse relation

## 5. Noise-robust AOIs for face stimuli

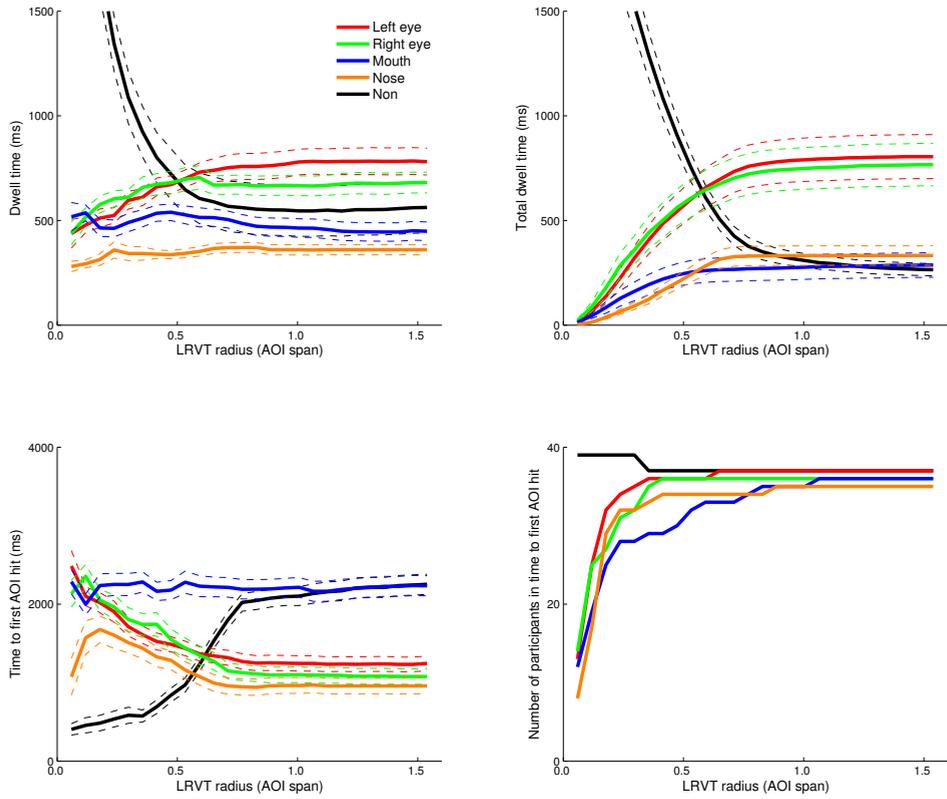


Figure 5.5.: Relation between LRVT radius (given in terms of AOI span) and mean dwell time (top left), mean total dwell time (top right), and mean time to the first AOI hit (bottom left) for the left eye, right eye, nose, mouth, and non AOI areas for the infant dataset. The numbers of participants for whom a time to first AOI hit could be calculated are depicted in the bottom right panel. Colored dashed lines indicate the standard errors of the means.

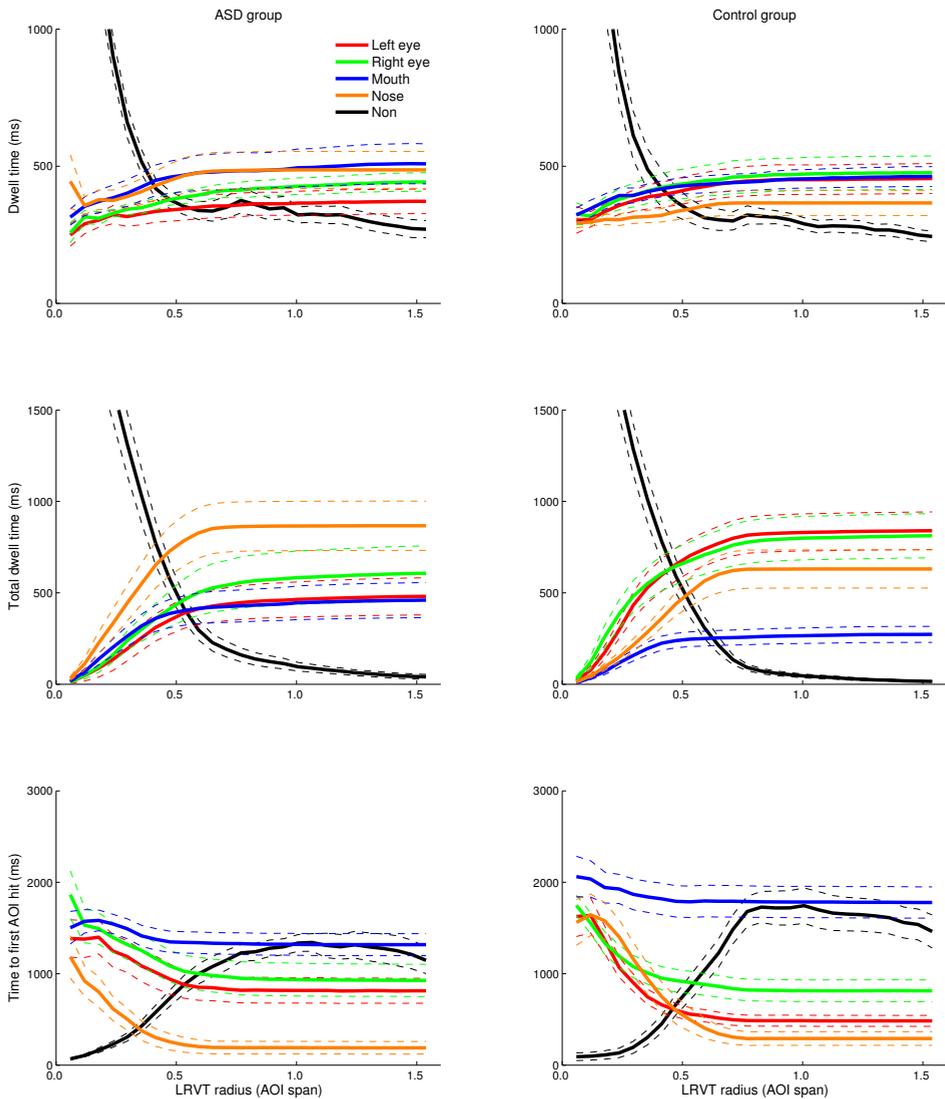


Figure 5.6.: Relation between LRVT radius (given in terms of AOI span) and mean dwell time (top panels), mean total dwell time (middle panels), and mean time to first AOI hit (bottom panels) for the left eye, right eye, nose, mouth, and non AOI areas. Panels in the left column are for the ASD group, and the panels in the right column for the control group. Colored dashed lines indicate the standard errors of the means.

## 5. Noise-robust AOIs for face stimuli

with LRVT radius, and increased up to an LRVT radius of roughly 0.75 AOI span ( $3^\circ$ ).

Following the asymptotic relation between LRVT radius and total dwell time we examined the effect of LRVT radius on statistical comparisons between AOI attention-maintaining capacities. This was done to examine whether AOI size matters for statistical outcomes. Paired-samples t-tests were carried out for the comparison of the mean total dwell time to the eye AOIs (sum of the total dwell time to left eye and right eye AOIs) and mouth AOI. This was done separately for the ASD and for the control group. As visible from Figure 5.7, the difference in total dwell time to the eyes and mouth for the ASD group was significant at  $\alpha = 0.05$  from an LRVT radius of 0.48 AOI span ( $1.9^\circ$ ) onwards. For the control group, however, the difference was significant from the smallest LRVT radius onwards. In addition, we examined the between-group comparison of mean total dwell time to the eye AOIs. As visible from Figure 5.8, the independent samples t-test on mean total dwell time to the eyes between the ASD and control group was significant at  $\alpha = 0.05$  between 0.18 and 0.53 AOI span ( $0.7^\circ$ - $2.1^\circ$ ) of LRVT radius. Hereafter, the p-value for the comparison increased slightly, but remained between 0.05 and 0.08 up to 1.5 AOI span of LRVT radius. We address the implications hereof for research comparing gaze behavior in ASD groups and typically developing controls in the discussion.

### 5.2.3. Question 3 – Which AOI type is most robust to noise?

We investigated the robustness of AOI types to noise by adding Gaussian noise with increasing standard deviation to fixation locations. Gaussian noise on the fixation location mimics a variable error in the data due to imprecision of the eye-tracking measurement. A range of 0.05 to 0.75 AOI span ( $0.2^\circ$  to  $3.0^\circ$ ) of Gaussian noise standard deviation was added. The reason for the cutoff of 0.75 ( $3^\circ$ ) is three-fold. First,  $3^\circ$  represents three times the size of noise inherent in the worst eye-trackers (see e.g. Holmqvist et al., 2011). Second, following the LRVT radius analysis (Question 2, see above), the differences in dwell times, total dwell time, and time to first AOI

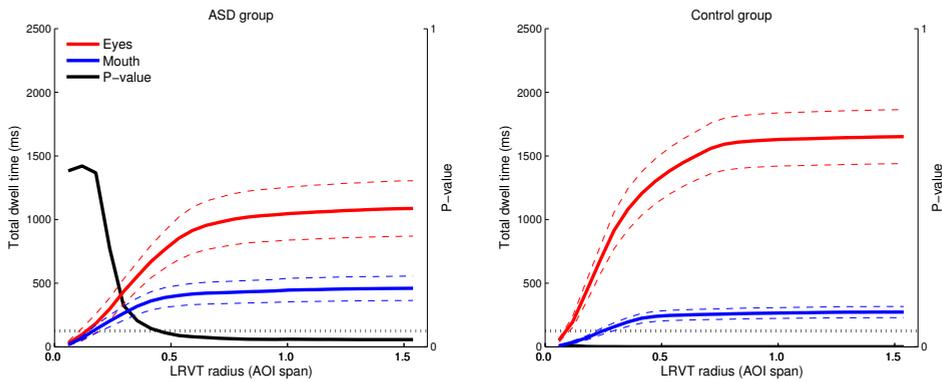


Figure 5.7.: Mean total dwell times to the combined eye AOIs and the mouth AOI for the ASD group (left) and the control group (right). Colored dashed lines indicate the standard errors of the means. The black lines depict the  $p$  values of the paired-samples  $t$  test carried out for each LRVT radius, with  $p = .05$  being indicated by the dotted horizontal line. Note that the  $p$ -value line for the control group is at 0 for all LRVT radii

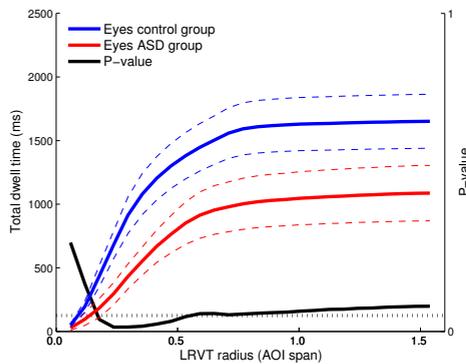


Figure 5.8.: Mean total dwell times to the eye AOIs for the ASD group and the control group. Colored dashed lines indicate the standard errors of the means. The black line depicts the  $p$  value of the independent-samples  $t$ -test carried out for each LRVT radius, with  $p = .05$  being indicated by the dotted horizontal line

## 5. Noise-robust AOIs for face stimuli

hit were most present between LRVT radii of 0 and 0.75 AOI span. Third, half the distance between the eyes in the stimulus is 0.75 AOI span. This means that if a fixation is positioned right between the eyes, a horizontal shift between minus and plus one standard deviation will result in a new location anywhere between the left eye and right eye.

Robustness to noise was operationalized as the slope of linear fit between the standard deviation of added Gaussian noise (in AOI span) and the group mean on the eye-tracking measure. A slope of 0 would indicate perfect robustness to noise: no increase or decrease in the group mean of dwell time, total dwell time, or time to first AOI hit. The further the slope is away from 0 (whether a positive or negative slope), the less robust to noise the AOI-production method is.

Figure 5.9 depicts the slopes for mean dwell time, mean total dwell time, and mean time to first AOI hit versus noise for the infant (left panels), ASD (middle panels) and control (right panels) participants. As visible from the top left panel in Figure 5.9, mean dwell times to the left eye, right eye and mouth AOIs tended to decrease as noise increased for most AOI-production methods. Mean dwell time for the nose AOI remained relatively stable for all production methods, whereas mean dwell time to the non AOI increased for all but the Voronoi AOI-production methods. The fact that dwell time to the non AOI did not increase for the Voronoi method was not surprising, as only data outside the screen were labeled as belonging to the non AOI. The slopes were closest to zero for all AOIs for the Voronoi method. As visible from the middle left panel in Figure 5.9, mean total dwell times decreased for the eye AOIs and the mouth AOI, for all but the Voronoi AOI-production method. The mean total dwell time to the non AOI increased for all but the Voronoi AOI-production method as noise increased. Again the absence of an increase in total dwell time to the non AOI for the Voronoi method was not surprising, given that only data outside the screen were labeled as belonging to the non AOI. As visible from the bottom left panel in Figure 5.9, confidence intervals of slopes for the mean time to first AOI hit were larger than for the (total) dwell times

for all AOI-production methods; indicating that the changes here were less consistent with increasing noise. This might be expected as mean time to first AOI hit was based only on the first fixation in the AOI in each trial. One shifted fixation early in the trial will already result in the measure for that trial being considerably lower. In general, time to first AOI hit for the nose increased with increasing noise, and time to first AOI hit for the non AOI decreased with increasing noise.

As visible from the middle and right top panels in Figure 5.9, mean dwell times to all feature AOIs were relatively stable across different noise levels for all AOI-production methods. Mean dwell time to the non AOI increased as noise increased for the Hand-drawn and LRVT methods for both groups. As visible from the middle panels for the ASD and control groups in Figure 5.9, mean total dwell times to the feature AOIs tended to decrease, while mean total dwell time to the non AOI increased with increasing noise. This was, however, much less the case for the Voronoi method in the ASD group, and altogether absent for the Voronoi method in the control group. In the latter case, the confidence intervals of all slopes were closely positioned around zero. As visible from the bottom panels for the ASD and control groups in Figure 5.9, confidence intervals for slopes for the mean time to first AOI hit were relatively large compared to the confidence intervals for slopes for mean dwell time and mean total dwell time. Mean time to first AOI hit tended to increase for the nose AOI, and decrease for the mouth and non AOI, as noise increased for most AOI-production methods. The slopes for the other AOIs were, however, less consistent over AOI-production methods.

The fact that the slopes were closest to zero for the Voronoi method, at least for the dwell time and total dwell time, led us to further investigate the effects of AOI size on robustness to noise. If the Voronoi method is most robust to noise purely because the AOIs are larger compared to the other AOI-production methods, slopes should approach zero with increasing AOI size. The same approach to AOI size as in question 2 of the results section was used to investigate this; i.e. by varying the radius of LRVT AOIs.

## 5. Noise-robust AOIs for face stimuli

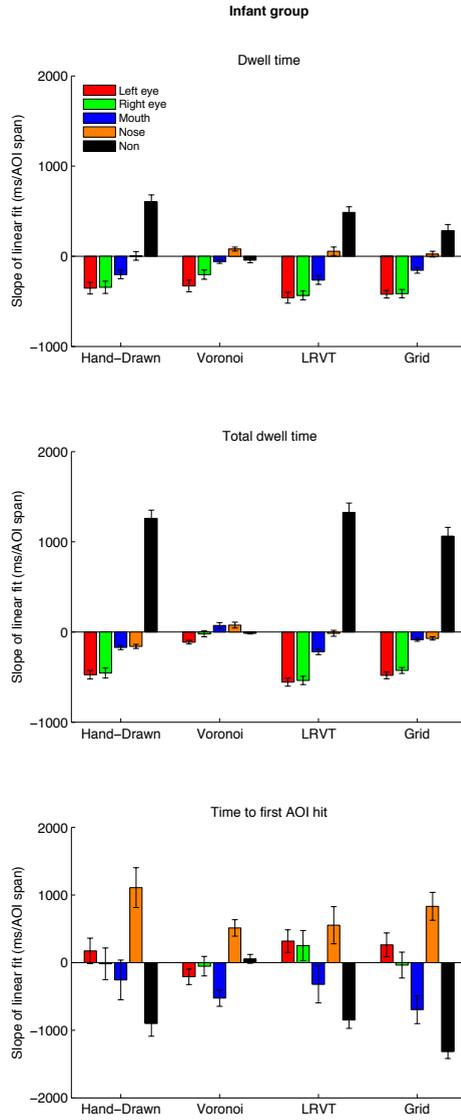
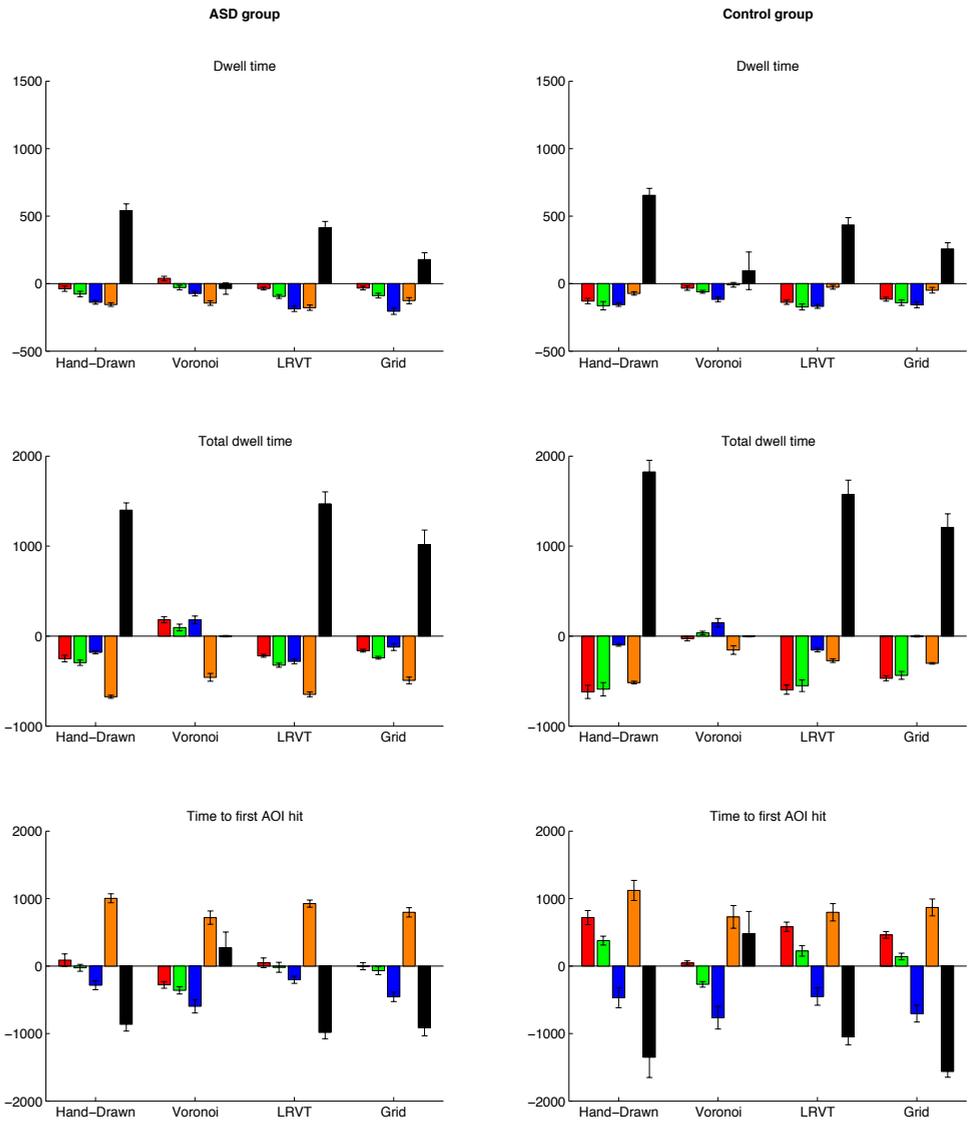


Figure 5.9.: Slopes for the linear fits between the standard deviations of Gaussian noise ( $\sigma$ ) and the dwell time (top panels), total dwell time (middle panels), and time to first AOI hit (bottom panels). Error bars indicate 95% confidence intervals of the slopes. Panels in the left column are for the infant group, panels in the middle column for the ASD group, and panels in the right column for the control group.

## 5.2. Results



## 5. Noise-robust AOIs for face stimuli

Figure 5.10 depicts the slopes for the linear fit between standard deviation of Gaussian noise and dwell time (top panels), total dwell time (bottom panels) as a function of LRVT radius. As visible from the top panels in Figure 5.10, slopes for dwell time were close to zero when LRVT radius was minimal – very little data was included in the AOIs at this point, see also Question 2 of the results section – but the absolute slope increased hereafter. Slopes for dwell time approached zero when LRVT radius increased beyond  $2^\circ$ , particularly for the infant group, but to a lesser extent for the ASD and control groups. As visible from the bottom panels in Figure 5.10, slopes for total dwell time were again close to zero when LRVT radius was minimal – very little data was included in the AOIs at this point – but the absolute slopes increased hereafter. The slopes for total dwell time approached zero when LRVT radius increased beyond  $2^\circ$  for all groups. Indeed, larger AOIs were more robust to noise, and especially for total dwell time.

### 5.3. Discussion

A problem in eye-tracking research is choosing Areas of Interest (AOIs): researchers in the same field often use widely varying AOIs for similar stimuli, making cross-study comparisons difficult or even impossible. Moreover, subjective choices mean that AOIs differ in shape, size, and location, and whether they are man-made or made using a computer. There are, however, not many guidelines for constructing AOIs or comparisons between AOI-production methods available. In the present study we addressed this by comparing AOI-production methods in face stimuli, using data collected with both infants and adults (ASD and matched controls). Specifically, we report that attention-attracting and attention-maintaining capacities of AOIs differ between AOI-production methods, and that this matters for statistical comparison in one of three groups investigated (the ASD group). In addition, we investigated the relation between AOI size and AOI attention-attracting and attention-maintaining capacity, and the consequences for statistical analyses, and report that adopting large AOIs solves the problem of statistical differences between AOI methods. Finally, we tested

AOI-production methods for their robustness to noise, and report that the Voronoi tessellation method is most robust to noise.

We first report that feature AOIs (eyes, nose, and mouth) may differ in size, location, and shape between AOI-production methods. However, attention-attracting and attention-maintaining capacity of AOIs across AOI-production methods appeared to show the same global pattern (e.g. longer total dwell times to the eyes compared to the nose for the infant or control group). When tested statistically, the differences between AOI-production methods were large enough, though, to affect the outcome for the ASD group. Using one of four AOI-production methods the difference in mean total dwell to the eyes was not significantly longer than to the mouth, whereas this was the case for the infant and control participants. As this is a particular relevant analysis in the ASD literature (Guillon et al., 2014), the finding that AOI-production method affected the outcome of statistical tests is not trivial. If the purpose of a study is to compare attention-attracting or attention-maintaining capacity between feature AOIs in ASD, it makes sense to justify the AOIs used in light of the present finding. If, on the other hand, the purpose of the study is to compare attention-attracting and attention-maintaining capacity between feature AOIs for infants or typically developing adults, it should not matter much which AOI-production method is used. It would then make sense to choose the AOI-production method that is most objective and easy to implement. In the present study, the most objective method would be the Voronoi method, of which only the cell centers are subjective. In addition, implementation of the Voronoi method is easy to do by machine once cell centers and fixations have been identified. If, on the other hand, the purpose of a study is to compare absolute values to feature AOIs to other studies, one should take care to construct AOIs using the same AOI-production method for each.

The relation between AOI size and AOI attention-attracting and attention-maintaining capacity was investigated using the LRV method. LRV radius was varied between 0.05 and 1.52 AOI span (see AOI span for details; corresponds to 0.2 - 6.0°). We report that for both the infant dataset,

## 5. Noise-robust AOIs for face stimuli

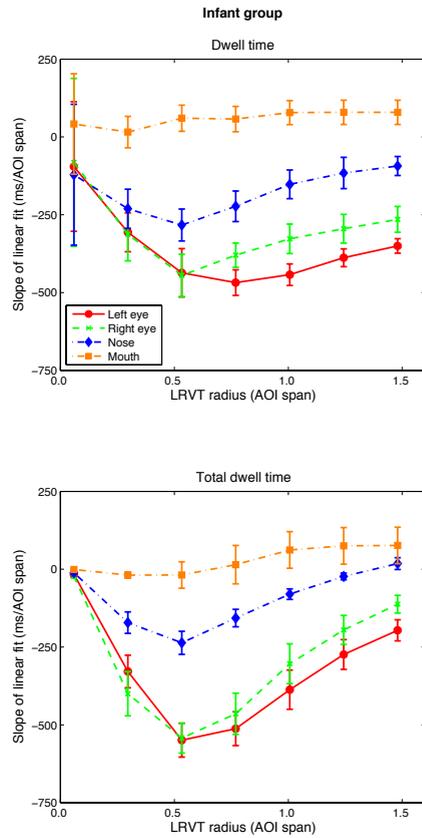
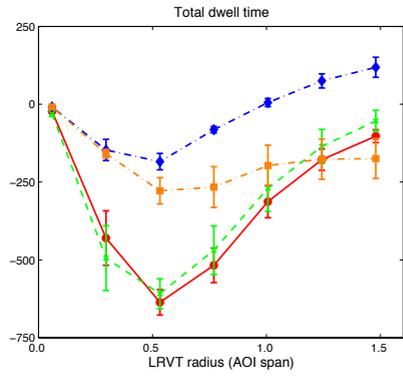
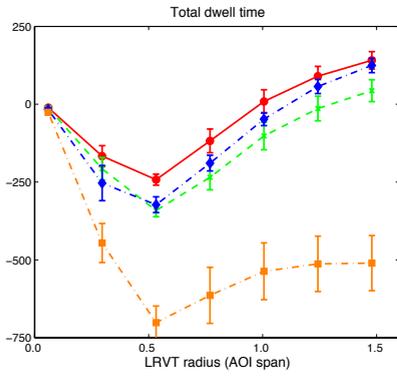
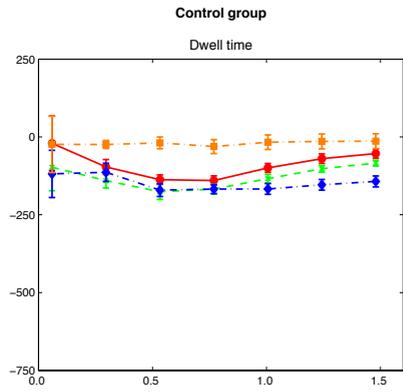
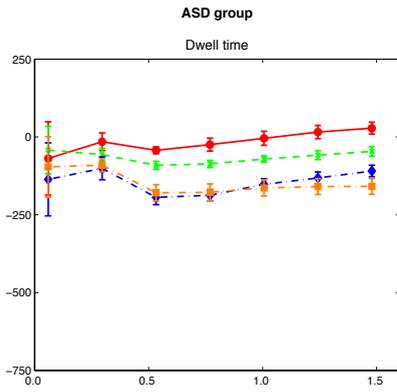


Figure 5.10.: Slopes for the linear fits between the standard deviations of Gaussian noise ( $^{\circ}$ ) and the dwell time (top panels) and total dwell time (bottom panels), as a function of LRVT radius (AOI span). Error bars indicate 95% confidence intervals of the slopes. Panels in the left column are for the infant group, panels in the middle column for the ASD group, and panels in the right column for the control group.



## 5. Noise-robust AOIs for face stimuli

and the ASD and control dataset, attention-maintaining capacity of feature AOIs increased as a function of LRVT radius, but approached an asymptote around an LRVT radius of 0.75 AOI span ( $3^\circ$ ). For attention-attracting capacity the same relation was demonstrated; as LRVT radius increased, so did attention-attracting capacity up to 0.75 AOI span ( $3^\circ$ ). Note that an increasing attention-attracting capacity is operationalized as a decreasing time to first AOI hit. For the infant dataset, attention-attracting capacity of feature AOIs was quite inconsistent when the LRVT radius was below 0.5 AOI span ( $2^\circ$ ). The reason hereof was that the number of participants that fixated the feature AOI within the small radius was low for small radii. As the radius increased, so did the number of participants for which time to first AOI hit could be calculated. In the adult dataset, where there were more trials, time to first AOI hit could be calculated for almost all participants from the smallest radius onwards. One possible solution for calculating attention-attracting capacity in situations where only a number of participants fixate a particular AOI is to use the  $T_{50}$  (Hooge & Camps, 2013). The  $T_{50}$  is the time it takes for 50% of the participants to hit a particular AOI, and can be calculated instead of a mean over all participants with varying number of participants at each data point. The idea behind the  $T_{50}$  is that AOIs with a high attention-attracting capacity are assumed to attract gaze more quickly, and for a higher number of participants, than AOIs with a low attention-attracting capacity.

The fact that total dwell times approached an asymptote around 0.75 AOI span ( $3^\circ$ ) LRVT radius is evidence for fixations being clustered on the stimulus, which is often the case for sparse stimuli. Increasing the size of the AOI no longer includes substantially more data, which means that gaze was directed mostly towards the limited AOI set. If a stimulus is sparse, increasing the size of the AOI would thus mean that more data are included, and differences between or within groups on total dwell time to feature AOIs might be more easily detectable using statistical models. This is also what we report here: if an LRVT radius below 0.48 AOI span is used, one might conclude that the ASD group does not look more towards the eyes than to the mouth, whereas for the control group that is the

case. If, on the other hand, an LRVT radius above 0.48 AOI span were used, one would conclude that for both the ASD and the control group, gaze is directed towards the eye region for a longer total period of time compared to the mouth region. Moreover, if no differences between groups were expected within the area not covered by feature AOIs – which in that case would warrant the construction of additional AOIs – it would also make sense to increase AOI size. If an LRVT radius below 0.18 AOI span were used, one would conclude that the ASD group does not look to the eye region for a shorter total period of time than the control group. If, however, an LRVT radius above 0.18 AOI span were used, one would conclude that the ASD group does, in fact, look to the eye region for a shorter total period of time than the control group. Providing arguments for AOI-production method and AOI size is recommended for future studies. This is particularly relevant for the literature on face scanning in ASD as there have been several inconsistent reports on whether individuals with ASD scan faces differently from typically developing controls (see Guillon et al., 2014, for a review). We suggest using large AOIs when making cross-group comparisons, such that as much data as possible is included.

Finally, the robustness of AOI-production methods to noise was investigated. If the mean values for attention-attracting and attention-maintaining capacity using a particular AOI-production method remain stable across increasing levels of Gaussian noise, this indicates that this method is robust to noise. We report here that the Voronoi method was the most noise-robust method in three participant groups from two different datasets when attention-maintaining capacity was considered. The reason the Voronoi method was most robust to noise is because its AOIs were largest. Systematically increasing the size of AOIs increases robustness to noise. Attention-maintaining capacity for all AOIs using the Voronoi method remained stable across increasing level of Gaussian noise with a maximum standard deviation of 0.75 AOI span ( $3^\circ$ ). This finding is particularly interesting given recent research. Wass, Forssman, & Leppänen (2014) added Gaussian noise to raw data samples from infant eye-tracking data to faces and report that the distribution of total dwell time across the eyes, nose, and mouth (hand-

## 5. Noise-robust AOIs for face stimuli

drawn rectangle shaped AOIs), and non AOI changed. They observed that increased noise is associated with a lower proportion of total dwell time to the eyes, and a higher proportion of total dwell time to the nose, mouth, and non AOI. They conclude that using non-contiguous AOIs might help reduce this error. If we compare their result to the data presented here, we see the same pattern of result for the slopes of mean total dwell time as a function of noise for the hand-drawn AOIs in the infant group. Mean total dwell time to the eyes decreased as a function of noise, decreased much less for the nose and mouth AOIs, and increased for the non AOI. If we translate this to a proportion of looking time, we would find a decrease in proportion of total dwell time to the eyes, and consequently a slightly increased proportion of total dwell time to the nose, mouth, and non AOI. When we consider the Voronoi AOI-production method we found almost no decrease or increase in mean total dwell time as a function of noise, particularly for the infant and control groups. We should note that the standard deviation of the Gaussian noise (i.e.  $3^\circ$ ) added in this analysis is much larger than the error typically reported by eye-tracker manufacturers (see e.g. Holmqvist et al., 2011). Under normal circumstances and when data quality checks are ensured, smaller errors due to noise may be expected than reported here. Consequently, we argue here that adopting large areas of interest – most easily implemented using Voronoi AOIs, optionally extended with a large radius as in the LRVT method – in faces might be a better solution to account for noisy data, for example, in infant eye-tracking research, although they should not serve as a replacement for acquiring high data quality (see e.g. Hessels, Andersson, Hooge, Nyström, & Kemner, 2015; Nyström et al., 2013).

Adopting large AOIs may seem counterintuitive given that Holmqvist et al. (2011) suggest that AOIs should be as precise as possible with regard to the objects of interest in the stimulus. However, as Hooge & Camps (2013) already point out, for relatively empty stimuli where there is not much crowding (lateral masking), such as faces, AOIs should be as large as possible. Here we show that large AOIs include all relevant data, and are most noise-robust. In addition, large AOIs constructed using the Voronoi

method are also easily implemented. While large AOIs are most suitable for hypothesis-driven research with a clear division of the relevant areas in a stimulus, fine-grained spatial effects cannot be uncovered using large AOIs. As pointed out above, when such fine-grained spatial effects are hypothesized, they require additional AOIs to be defined as opposed to only the main feature AOIs presented here. Moreover, when such fine-grained spatial effects can not be hypothesized in terms of easily distinguishable AOIs, but are of the researcher's interest, data-driven approaches that statistically compare entire fixation maps may be much better suited (see e.g. Caldara & Miellet, 2011). Such data-driven approaches have already been adopted for face processing in adults (Arizpe, Kravitz, Yovel, & Baker, 2012; Blais, Jack, Scheepers, Fiset, & Caldara, 2008; Caldara, Zhou, & Miellet, 2010), infants (W. S. Xiao, Xiao, Quinn, Anzures, & Lee, 2013), and in ASD (Shi et al., 2015; Yi et al., 2013; 2015). Concluding, the purpose of the research, combined with the stimuli used, should drive the choice for AOI- or analysis-type.

## 5.4. Conclusions and limitations

We report here that the attention-attracting and attention-maintaining capacity of feature AOIs in faces relative to one another do not differ drastically between AOI-production methods. In addition, we conclude that large AOIs using the Voronoi method or LRVT with large radii are the most objective of the researcher-defined AOIs and the most noise-robust AOI-production method for use in face stimuli. It is particularly appealing as it can be implemented using a simple computer script, requiring only coordinates of AOI cell centers and fixation locations.

We reason here that, as faces are sparse stimuli, adopting larger AOIs using the Voronoi or LRVT method might generally be the preferred method in sparse stimuli. Other types of sparse stimuli were, however, not investigated in the present study and we therefore suggest this advice should be taken with caution. We welcome future research into the effects of AOI-production methods in other sparse stimuli as well as dense stimuli, where

## 5. *Noise-robust AOIs for face stimuli*

we expect other AOI-production methods to thrive.

### **Acknowledgements**

The authors would like to thank Maartje de Jong for creating the stimuli, Aliegriet Pol for help with data collection, Esther Eijlers for creating the hand-drawn AOIs, and Siarhei Uzunbajakau for building the experimental setup. This work was supported by a Netherlands Organization for Scientific Research (NWO) VICI grant (45307004) and the Consortium on Individual Development (CID). CID is funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the NWO (grant number 024.001.003).

## References

- Arizpe, J., Kravitz, D. J., Yovel, G., & Baker, C. I. (2012). Start position strongly influences fixation patterns during face processing: Difficulties with eye movements as a measure of information use. *PLOS One*, 7(2):e31106.
- Aurenhammer, F. (1991). Voronoi diagrams—a survey of a fundamental geometric data structure. *ACM Computing Surveys*, 23(3):345–405.
- Blais, C., Jack, R. E., Scheepers, C., Fiset, D., & Caldara, R. (2008). Culture shapes how we look at faces. *PLOS One*, 3(8):e3022.
- Caldara, R., & Miellet, S. (2011). iMap: A novel method for statistical fixation mapping of eye movement data. *Behavior Research Methods*, 43(3):864–878.
- Caldara, R., Zhou, X., & Miellet, S. (2010). Putting culture under the ‘spotlight’ reveals universal information use for face recognition. *PLOS One*, 5(3):e9708.
- Chawarska, K., & Shic, F. (2009). Looking but not seeing: Atypical visual scanning and recognition of faces in 2 and 4-year-old children with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, 39(12):1663–1672.
- de Jong, M. C., van Engeland, H., & Kemner, C. (2008). Attentional effects of gaze shifts are influenced by emotion and spatial frequency, but not in autism. *Journal of the American Academy of Child & Adolescent Psychiatry*, 47(4):443–454.
- Falck-Ytter, T. (2008). Face inversion effects in autism: a combined looking time and pupillometric study. *Autism Research*, 1(5):297–306.
- Gallay, M., Baudouin, J.-Y., Durand, K., Lemoine, C., & Lécuyer, R. (2006). Qualitative differences in the exploration of upright and upside-down faces in four-month-old infants: An eye-movement study. *Child Development*, 77(4):984–996.
- Goldberg, J. H., & Helfman, J. I. (2010). Comparing information graphics: A critical look at eye tracking (pp. 1–8). *Proceedings of the 3rd BELIV’10 workshop: Beyond time and errors: novel evaluation methods for information visualization*.
- Guillon, Q., Hadjikhani, N., Baduel, S., & Rogé, B. (2014). Visual social attention in autism spectrum disorder: Insights from eye tracking studies. *Neuroscience & Biobehavioral Reviews*, 42:279–297.
- Hessels, R. S., Andersson, R., Hooge, I. T. C., Nyström, M., & Kemner, C. (2015). Consequences of eye color, positioning, and head movement for eye-tracking data quality in infant research. *Infancy*, 20(6):601–633.
- Holmqvist, K., Nyström, M., & Mulvey, F. (2012). Eye tracker data quality: What

## 5. Noise-robust AOIs for face stimuli

- it is and how to measure it. *Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '12*, 45.
- Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press.
- Hooge, I., & Camps, G. (2013). Scan path entropy and arrow plots: capturing scanning behavior of multiple observers. *Frontiers in Psychology*, 4:996.
- Hunnus, S., & Geuze, R. H. (2004). Developmental changes in visual scanning of dynamic faces and abstract stimuli in infants: A longitudinal study. *Infancy*, 6(2):231–255.
- Hunnus, S., de Wit, T. C. J., Vriens, S., & Hofsten, von, C. (2011). Facing threat: Infants' and adults' visual scanning of faces with neutral, happy, sad, angry, and fearful emotional expressions. *Cognition & Emotion*, 25(2):193–205.
- Jones, W., & Klin, A. (2013). Attention to eyes is present but in decline in 2–6-month-old infants later diagnosed with autism. *Nature*, 504:427–431.
- Jones, W., Carr, K., & Klin, A. (2008). Absence of preferential looking to the eyes of approaching adults predicts level of social disability in 2-year-old toddlers with autism spectrum disorder. *Archives of General Psychiatry*, 65(8):946–954.
- Kano, F., & Tomonaga, M. (2010). Face scanning in chimpanzees and humans: continuity and discontinuity. *Animal behaviour*, 79:227–235.
- Liu, S., Quinn, P. C., Wheeler, A., Xiao, N., Ge, L., & Lee, K. (2011). Similarity and difference in the processing of same- and other-race faces as revealed by eye tracking in 4- to 9-month-olds. *Journal of Experimental Child Psychology*, 108:180–189.
- Nguyen, H. T., Isaacowitz, D. M., & Rubin, P. A. D. (2009). Age- and fatigue-related markers of human faces: An eye-tracking study. *Ophthalmology*, 116(2):355–360.
- Nyström, M., Andersson, R., Holmqvist, K., & van de Weijer, J. (2013). The influence of calibration method and eye physiology on eyetracking data quality. *Behavior Research Methods*, 45(1):272–288.
- Oakes, L. M., & Ellis, A. E. (2013). An eye-tracking investigation of developmental changes in infants' exploration of upright and inverted human faces. *Infancy*, 18(1):134–148.
- Over, E. A. B., Hooge, I. T. C., & Erkelens, C. J. (2006). A quantitative measure for the uniformity of fixation density: The Voronoi method. *Behavior Research Methods*, 38(2):251–261.
- Rutherford, M. D., & Towns, A. M. (2008). Scan path differences and similari-

- ties during emotion perception in those with and without autism spectrum disorders. *Journal of Autism and Developmental Disorders*, 38(7):1371–1381.
- Senju, A., Verneti, A., Kikuchi, Y., Akechi, H., & Hasegawa, T. (2013). Cultural modulation of face and gaze scanning in young children. *PLOS One*, 8(8):e74017.
- Shi, L., Zhou, Y., Ou, J., Gong, J., Wang, S., Cui, X., Lyu, H., Zhao, J., & Luo, X. (2015). Different visual preference patterns in response to simple and complex dynamic social stimuli in preschool-aged children with autism spectrum disorders. *PLOS One*, 10(3):e0122280.
- Shic, F., Macari, S., & Chawarska, K. (2014). Speech disturbs face scanning in 6-month-old infants who develop autism spectrum disorder. *Biological Psychiatry*, 75(3):231–237.
- Tenenbaum, E. J., Shah, R. J., Sobel, D. M., & Malle, B. F. (2013). Increased focus on the mouth among infants in the first year of life: A longitudinal eye-tracking study. *Infancy*, 18(4):534–553.
- van Casteren, M., & Davis, M. H. (2006). Mix, a program for pseudorandomization. *Behavior Research Methods*, 38(4):584–589.
- Võ, M. L. H., Smith, T. J., Mital, P. K., & Henderson, J. M. (2012). Do the eyes really have it? Dynamic allocation of attention when viewing moving faces. *Journal of Vision*, 12(13):3.
- Voronoi, G. (1909). Nouvelles applications des paramètres continus à la théorie des formes quadratiques: Deuxième mémoire. Recherches sur les paralléloèdres primitifs. *Journal für die Reine & Angewandte Mathematik*, 136:67–182.
- Wagner, J. B., Luyster, R. J., Yim, J. Y., Tager-Flusberg, H., & Nelson, C. A. (2013). The role of early visual attention in social development. *International Journal of Behavioral Development*, 37:118–124.
- Wass, S. V., Forssman, L., & Leppänen, J. (2014). Robustness and precision: How data quality may influence key dependent variables in infant eye-tracker analyses. *Infancy*, 19(5):427–460.
- Wheeler, A., Anzures, G., Quinn, P. C., Pascalis, O., Omrin, D. S., & Lee, K. (2011). Caucasian infants scan own- and other-race faces differently. *PLOS One*, 6(4):e18621.
- Wilcox, T., Stubbs, J. A., Wheeler, L., & Alexander, G. M. (2013). Infants' scanning of dynamic faces during the first year. *Infant Behavior and Development*, 36(4):513–516.
- Xiao, W. S., Xiao, N. G., Quinn, P. C., Anzures, G., & Lee, K. (2013). Development of face scanning for own- and other-race faces in infancy. *International Journal of Behavioral Development*, 37(2):100–105.

## 5. Noise-robust AOIs for face stimuli

- Yi, L., Fan, Y., Quinn, P. C., Feng, C., Huang, D., Li, J., Mao, G., & Lee, K. (2013). Abnormality in face scanning by children with autism spectrum disorder is limited to the eye region: Evidence from multi-method analyses of eye tracking data. *Journal of Vision*, 13(10):5.
- Yi, L., Quinn, P. C., Feng, C., Li, J., Ding, H., & Lee, K. (2015). Do individuals with autism spectrum disorder process own- and other-race faces differently? *Vision Research*, 107:124–132.

**Part II.**

# **Visual search in ASD**



## **6. Is there a limit to the superiority of individuals with ASD in visual search?**

Published as:

Hessels, R. S., Hooge, I. T. C., Snijders, T. M., & Kemner, C. (2014). Is there a limit to the superiority of individuals with ASD in visual search? *Journal of Autism and Developmental Disorders*, 44(2):443–451.

Author contributions:

IH, TS, CK designed the study. TS collected the data. RH, IH analyzed the data. RH, CK, IH interpreted the data. RH drafted the paper. RH, IH, TS, CK finalized the paper.

## **Abstract**

Superiority in visual search for individuals diagnosed with autism spectrum disorder (ASD) is a well-reported finding. We administered two visual search tasks to individuals with ASD and matched controls. One showed no difference between the groups, and one did show the expected superior performance for individuals with ASD. These results offer an explanation, formulated in terms of load theory. We suggest that there is a limit to the superiority in visual search for individuals with ASD, related to the perceptual load of the stimuli. When perceptual load becomes so high that no additional task-(ir)relevant information can be processed, performance will be based on single stimulus identification, in which no differences between individuals with ASD and controls have been demonstrated.

Individuals who have been diagnosed with Autism Spectrum Disorder (ASD) typically exhibit impairments in social interaction and communication, whilst on the other hand excel on certain visuospatial tasks (see Dakin & Frith, 2005, and Simmons et al., 2009, for an overview). This is possibly highly beneficial for early recognition of ASD, and indeed, studies on processing of eye-gaze and visual fixation during social interaction in infancy have already been successful in distinguishing infants at-risk for ASD from controls (Elsabbagh et al., 2009; Merin, Young, Ozonoff, & Rogers, 2007). Non-social visuospatial differences might prove to be the next step in early recognition of ASD. However, such prospects are only viable if the perceptual differences between individuals with ASD and typically developing individuals are well understood.

A particularly well-documented finding is the excellence of individuals with ASD in visual search (Jarrold, Gilchrist, & Bender, 2005; Joseph, Keehn, Connolly, Wolfe, & Horowitz, 2009; Kemner, Ewijk, Engeland, & Hooge, 2008; O’Riordan, Plaisted, Driver, & Baron-Cohen, 2001; O’Riordan, 2004; Plaisted, O’Riordan, & Baron-Cohen, 1998a), where individuals with ASD show shorter reaction times in visual search tasks when compared to a matched control group. This excellence manifests itself even as early as in 2,5-year old toddlers (Kaldy, Kraper, Carter, & Blaser, 2011). In this study, eye-movements of toddlers were measured and toddlers with ASD proved more successful at finding a target distinct by a combination of features (Kaldy et al., 2011). A possible explanation for these findings is an enhanced stimulus discriminability in individuals with ASD, which is demonstrated by increasing differences between groups in reaction time when the task becomes more difficult (O’Riordan et al., 2001; Plaisted, O’Riordan, & Baron-Cohen, 1998a). This is also supported by the finding that children with ASD outperform a control group in an embedded figures tasks where the target is embedded in a more complex image (Jarrold et al., 2005). This superiority in visual search also seems to generalize to the typically developing population, where individuals high on the autistic traits spectrum, as measured with the Autism Spectrum Quotient (Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001), outperform indi-

## 6. *Limit to search superiority in ASD*

viduals low on the autistic traits spectrum in a visual search task (Almeida, Dickinson, Maybery, Badcock, & Badcock, 2010).

Another explanation might be that individuals with ASD adopt a different search strategy, and are more successful or faster because of this. To test whether different search strategies between individuals with ASD and controls might explain the enhanced performance of individuals with ASD, several studies looked at eye-movements during visual search (Joseph et al., 2009; Kemner et al., 2008). Kemner et al. (2008) conclude that individuals with ASD do not adopt a more efficient search strategy, which would be accompanied by longer fixation times (Hooge & Erkelens, 1999), and that the superior performance might thus be due to enhanced stimulus discrimination. In addition, Joseph et al. (2009) conclude that participants with ASD search the displays in a similar fashion as controls, and that the superior performance is due to non-search processes, i.e. stimulus discrimination, not search-strategy. Furthermore, children with ASD do not appear to search efficiently in a more ecological search setting (Pellicano et al., 2011).

The literature seems to provide a general consensus on superior performance for individuals with ASD in visual search, likely due to enhanced stimulus discrimination, which becomes more pronounced when task difficulty increases. This superior performance is evident from reduced reaction times in individuals with ASD, while error rates are equal to control participants. As this superior performance is well replicated, the present study aimed to use this as a baseline for comparison with effects between individuals with ASD and controls in parallel studies. A visual search task using a vertical target among tilted distractors was used to accomplish this. However, in this experiment we did not observe the expected visual search superiority of individuals with ASD compared to the control group. This incongruence with the literature led us to conduct a second experiment in which visual search superiority for individuals with ASD was again demonstrated. The results will be discussed and put in perspective with current theories on enhanced stimulus discrimination in ASD.

Table 6.1.: Means and standard deviations for age, IQ, AQ, contrast sensitivity and visual acuity for the ASD and the matched control group.

	Control group	ASD group
Sample size	31	19
Age (years)	21.6 (2.11)	21.6 (3.04)
IQ	118.9 (11.1)	116.2 (12.2)
AQ	24.4 (3.22)	32.7 (4.16)
Weber contrast sensitivity	1.69% (1.32%)	1.33% (0.58%)
Visual acuity (LogMAR)	-0.24 (0.16)	-0.25 (0.11)

## 6.1. Methods Experiment 1

### 6.1.1. Participants

19 young adults with ASD (4 female, 15 male) and 31 matched control subjects (7 female, 24 male) participated in experiment 1. Mean age of the ASD group was 21.6 years ( $sd = 3.04$  years). Mean age of the control group was 21.6 years ( $sd = 2.11$  years). For the ASD group, the Wechsler Adult Intelligence Scale III, Dutch edition (WAIS-III), was used to determine IQ scores. For the control group the Wechsler Abbreviated Scale of Intelligence (WASI) was used to estimate IQ. The Freiburg Visual Acuity Test (Bach, 1996) was used to measure visual acuity and Weber contrast sensitivity (i.e. sensitivity for contrast between feature and background luminance). The ASD and control groups did not differ significantly on age, IQ, visual acuity or contrast sensitivity, and all subjects had normal or corrected-to-normal vision (log MAR acuity range -0.49 to 0.35). The Autism Spectrum Quotient (AQ; Baron-Cohen et al., 2001) was administered to all subjects. Due to technical reasons AQ data of two control subjects was lost. Mean AQ for the patient group was 32.7 ( $sd = 4.16$ ) and 24.4 ( $sd = 3.22$ ) for the control group. An independent-samples t-test was used to determine significance,  $t(46) = 7.86$ ,  $p < .001$ . Descriptive statistics for both groups are given in Table 6.1.

## 6. *Limit to search superiority in ASD*

The diagnostic evaluation for the ASD group included a psychiatric observation and a review of prior records (developmental history, child psychiatric and psychological observations and tests). ASD was diagnosed by a child psychiatrist using the DSM-IV criteria. The ASD group consisted of 12 participants diagnosed with Asperger's Syndrome, 6 diagnosed with Autism, and 1 diagnosed with PDD-NOS. The parents of all but one of these subjects were administered the Autism Diagnostic Interview (Lord, Rutter, & Couteur 1994) and eleven of the participants with ASD were administered the Autism Diagnostic Observation Schedule-generic (Lord et al., 1989), both by a trained rater. Seventeen subjects met ADI-R criteria for autism or autism spectrum disorder; one subject did not (this subject did, however, meet DSM-IV criteria). All of the participants who completed the ADOS-G met the full criteria for autism or autism spectrum disorder.

Control subjects were recruited on campus and were screened for ASD, ASD in their family, and history of psychopathology. None of the control subjects reported ASD or psychopathology, although two subjects had first-degree relatives diagnosed with ASD. Both the subjects with ASD and the control participants received a money reward for their participation. The study was approved by the medical ethics committee of the University Medical Centre Utrecht and subjects gave written informed consent prior to participation.

### **6.1.2. Stimuli and task**

Search displays consisted of a dark-grey background containing 25 light-grey line elements (see Figure 6.1). In the 'absent' condition, all lines were tilted 10° clockwise. In the 'present' condition one line element (i.e. the target) was vertical, whilst the remaining 24 lines were again tilted 10° clockwise. 50 'absent' and 50 'present' trials were presented in a pseudorandom order. Participants performed a binary forced choice task indicating whether the target was absent or present. Each trial started with a fixation cross. After the subject pressed the space bar, the search display

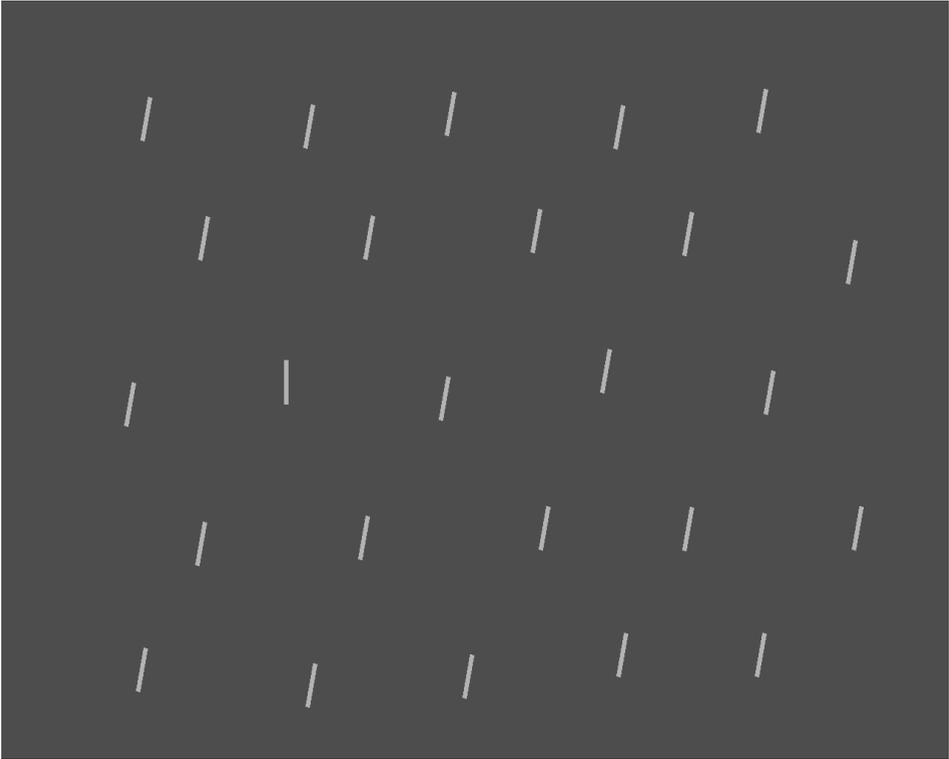


Figure 6.1.: Example stimulus in target 'present' condition.

appeared with a stimulus-onset asynchrony of 0-1000 ms. The search display remained on screen until the subject pressed one of the response keys. Response keys were the left and the right arrow-keys on the keyboard. The keys representing target absent and target present were counterbalanced across participants. Response on each trial initiated the fixation phase of the next trial.

### 6.1.3. Procedure

Participants were seated in a height-adjustable chair at 57 cm distance from a 22" monitor. A chin-rest was used to maintain viewing distance. The size of the search displays was  $31.25^\circ \times 25^\circ$ . A black curtain was then drawn over the participant and the screen so that no external light

## 6. Limit to search superiority in ASD

influenced the setup.

### 6.2. Results experiment 1

The mean reaction times and error rates were calculated for target present and target absent trials in both groups. One of the clinical subjects was removed from the analysis, as the mean reaction time for target absent trials for this participant was 10.3 s compared to a mean reaction time of the other young ASD adults of 2.77 s ( $sd = 1.03$  s). Other than that, no trials were excluded from the analysis.

#### 6.2.1. RT analysis

A repeated measures analysis of variance (ANOVA) with target (i.e. present or absent) as a within-subjects factor and group (i.e. ASD or control) as a between-subjects factor was used to determine statistical significance. The ANOVA revealed a significant effect of target on reaction time ( $F(1,47) = 89.22$ ,  $p < .0005$ ,  $\eta_p^2 = .66$ ), reflecting the fact that both groups were faster on target present trials as compared to target absent trials. No differences were observed between the ASD group and controls (see Figure 6.2). There was no effect of group ( $F < 1$ ), nor was there an interaction between target and group ( $F < 1$ ). The results are depicted in Figure 6.2.

#### 6.2.2. Error analysis

A repeated measures ANOVA revealed a significant effect of target on error rate ( $F(1,47) = 49.13$ ,  $p < .0005$ ,  $\eta_p^2 = .51$ ), indicating that the error rate in the target present condition ( $m = .142$ ,  $sd = .125$ ) was significantly higher than in the target absent condition ( $m = .007$ ,  $sd = .013$ ). Again, there was no effect of group ( $F < 1$ ), nor was there an interaction between target and group ( $F < 1$ ).

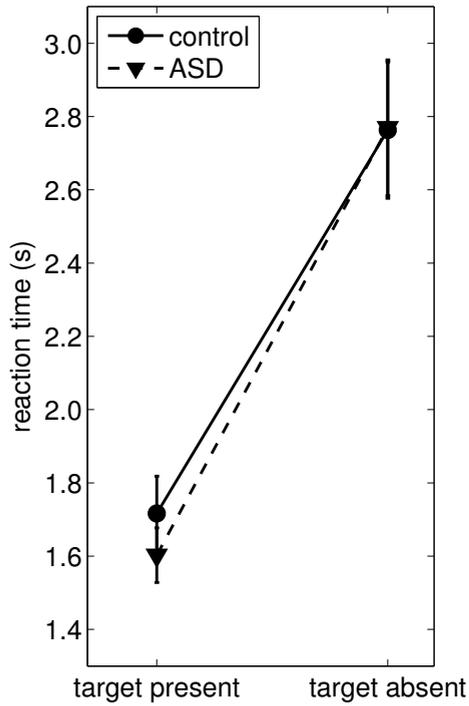


Figure 6.2.: Mean reaction times with SE for target absent and target present trials in ASD and controls.

### **6.3. Discussion experiment 1**

The purpose of experiment 1 was to replicate the general finding that the ASD group shows superior performance in visual search. To accomplish this, a visual search task was administered in which participants searched for a vertical target among tilted distractors. Both the ASD group and the control group were significantly slower to react when the target was absent, consistent with general findings in visual search. There was, however, no superior performance for the ASD group, neither when the target was present nor absent. More errors were made by both groups when the target was present compared to when the target was absent. However, no differences in error rates between the ASD group and the control group were identified, neither when the target was present nor when it was absent. These results are intriguing, and in high contrast with the body of research on visual search in ASD. The study by Kemner et al. (2008) used similar stimuli and a similar task as the present study, but, in contrast to the present study, did report clear differences in visual search performance between subjects with ASD and controls. The only differences between the two experiments were that the present experiment consisted of only one set size (i.e. 25 elements), and that the distractors were slightly less tilted from vertical (i.e.  $10^\circ$  clockwise tilt in the present experiment compared to a  $17^\circ$  clockwise tilt in Kemner et al. 2008). Furthermore, the groups in both studies were very similar regarding age, IQ and diagnosis. Finally, it should be noted that in the earlier study by Kemner et al. (2008), the effect between the ASD group and controls was numerically largest in the display resembling the present experiment the most; the 25-element display with a vertical target between  $17^\circ$  tilted distractors. To resolve this apparent incongruence, a second experiment was conducted with stimuli constructed to match those in Kemner et al. (2008).

Table 6.2.: Means and standard deviations for age, IQ, AQ, contrast sensitivity and visual acuity for the ASD and matched control group.

	Control group	ASD group
Sample size	14	13
Age (years)	22.8 (2.46)	22.9 (4.05)
IQ	118.9 (13.5)	116.1 (11.6)
AQ	24.7 (3.58)	32.4 (4.08)
Weber contrast sensitivity	1.27% (0.34%)	1.54% (0.42%)
Visual acuity (log MAR)	-0.26 (0.12)	-0.25 (0.11)

## 6.4. Methods experiment 2

### 6.4.1. Participants

13 Young adults with ASD and 14 matched control subjects participated in experiment 2. 12 control participants and 12 participants with ASD had previously participated in experiment 1. AQ scores for these participants were compared using an independent-samples t-test ( $t(22) = 4.95$ ,  $p < .001$ ). At least six months had passed between experiment 1 and 2. For the 2 control participants who had not participated in experiment 1, the Wechsler Abbreviated Scale of Intelligence was used to estimate IQ. To the young ASD adult who had not participated before the full Wechsler Adult Intelligence Scale was administered. This participant was diagnosed with Autism according to DSM-IV criteria, resulting in a group total of 7 diagnosed with Autism. Furthermore, the group comprised 5 participants diagnosed with Asperger’s Syndrome and 1 participant diagnosed with PDD-NOS. The ASD group and the matched control group did not differ significantly on age, IQ, visual acuity or contrast sensitivity. Descriptive statistics are provided in Table 6.2.

### 6.4.2. Stimuli and task

The experiment consisted of two sessions, one ‘easy’ and one ‘hard’ session (i.e. analogous to Kemner et al. 2008), which were counterbalanced across participants. Stimuli again consisted of dark-grey displays contain-

## 6. *Limit to search superiority in ASD*

ing light-grey line elements. The set-size of lines in the stimulus display was varied in both sessions with 4, 16 or 25 lines in the display. In the ‘easy’ session, the display consisted of vertical line distractors with a target line tilted 17° clockwise in half of the displays. In the ‘hard’ sessions, the display consisted of 17° clockwise tilted line distractors, with a vertical line target in half of the displays. 30 trials were presented per condition, resulting in a total of 180 trials (30 trials \* 2 (target present or absent) \* 3 (set-size)) per session. Participants performed a binary forced choice task indicating whether the target was absent or present. Each trial started with a fixation cross. After the subject pressed the space bar, the search display appeared immediately. The search display remained on screen until the subject pressed one of the response keys. Participants responded with the left and the right arrow-keys on the keyboard. The left arrow key represented the vertical line, and was to be pressed when the target was present in the ‘hard’ session and when the target was absent in the ‘easy’ session. The right arrow key represented the tilted line and was to be pressed when the target was present in the ‘easy’ session and when the target was absent in the ‘hard’ session. The response on each trial initiated the fixation phase of the next trial.

### 6.4.3. Procedure

The procedure in experiment 2 was identical to the procedure in experiment 1.

## 6.5. Results experiment 2

The mean reaction times and error rates were calculated for target absent and target present trials in all set sizes. Analyses were done for the ‘easy’ and ‘hard’ session separately. No trials were excluded from the analysis.

### 6.5.1. RT analysis

A repeated-measures ANOVA with Greenhouse-Geisser correction using target (absent or present) and set size (4, 16 or 25) as within-subjects

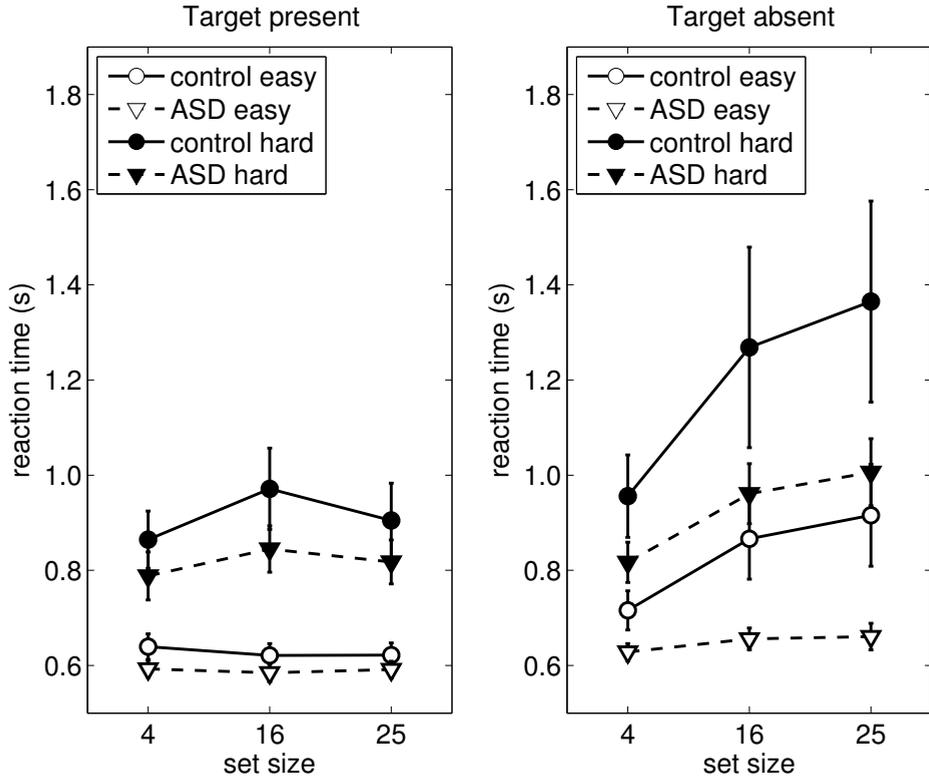


Figure 6.3.: Mean reaction times with SE. The left panel depicts reaction time at set sizes 4, 16 and 25 for ASD and control groups when the target was present. The right panel depicts reaction time when the target was absent.

factors and group (ASD or control) as a between-subjects factor was used to determine statistical significance. The results are depicted in Figure 6.3.

In the ‘easy’ session, a main effect for target was observed ( $F(1,25) = 18.96$ ,  $p < .0005$ ,  $\eta_p^2 = .43$ ), reflecting longer reaction times when the target was absent (609 ms for target present vs. 740 ms for target absent). Reaction times increased with set size ( $F(1.29,32.21) = 7.50$ ,  $p < .01$ ,  $\eta_p^2 = .23$ ) with mean reaction times of 644 ms, 682 ms and 697 ms, respectively. A significant two-way interaction between target and set size ( $F(1.12,28.04)$

## 6. Limit to search superiority in ASD

= 9.01,  $p < .005$ ,  $\eta_p^2 = .27$ ) reflected the rise in reaction time due to set size especially when the target was absent.

As can be seen in Figure 6.3, clear differences between the ASD and control group were found. The between-subjects effect for group was significant ( $F(1,25) = 4.25$ ,  $p = .05$ ,  $\eta_p^2 = .15$ ), indicating the overall reduction in reaction time for the ASD group compared to the control group. The two-way interaction between target and group ( $F(1,25) = 5.86$ ,  $p < .05$ ,  $\eta_p^2 = .19$ ) reflected the difference between groups especially for the target absent condition. Finally, the three-way interaction of target, set size and group was significant ( $F(1.12,28.04) = 4.53$ ,  $p < .05$ ,  $\eta_p^2 = .15$ ) and reflected the increase in reaction times due to set size when the target was absent, particularly for the control group.

In the ‘hard’ session, again longer reaction times were found for target absent trials ( $F(1,25) = 12.74$ ,  $p < .005$ ,  $\eta_p^2 = .34$ ); 865 ms for target present vs. 1062 ms for target absent. Reaction times increased with set size ( $F(1.08,27.03) = 8.33$ ,  $p < .01$ ,  $\eta_p^2 = .25$ ); with mean reaction times of 856 ms, 1011 ms and 1023 ms, respectively. A significant target and set size interaction was observed ( $F(1.39,34.86) = 12.02$ ,  $p < .0005$ ,  $\eta_p^2 = .33$ ). Although the pattern of results in the ‘hard’ session resembled that in the ‘easy’ session (see Figure 6.3), the main between-subjects effect for group ( $F(1,25) = 2.04$ ,  $p = .17$ ,  $\eta_p^2 = .08$ ), the target and group interaction ( $F(1,25) = 2.42$ ,  $p = .13$ ,  $\eta_p^2 = .09$ ), the set size and group interaction ( $F(1.08,27.03) = 1.01$ ,  $p = .33$ ,  $\eta_p^2 = .04$ ) and the target, set size and group three-way interaction ( $F(1.39,34.86) = 1.88$ ,  $p = .18$ ,  $\eta_p^2 = .07$ ) were not significant for the ‘hard’ condition.

### 6.5.2. Error analysis

Mean error rates in the ‘easy’ session were .025 ( $sd = .007$ ) in the target present condition and .016 ( $sd = .003$ ) in the target absent condition. A repeated-measures ANOVA on error rates in the ‘easy’ session revealed that these means did not differ significantly, nor were any other significant main

or interaction effects observed.

A repeated-measures ANOVA on error rates in the ‘hard’ session revealed a significant main effect for target ( $F(1,25) = 10.23$ ,  $p < .005$ ), indicating that the error rate in the target present condition ( $m = .044$ ,  $sd = .008$ ) was significantly higher than in the target absent condition ( $m = .015$ ,  $sd = .003$ ). No main between-subjects effect for group was observed, nor were there interactions involving the group factor.

## 6.6. General Discussion

The aim of present study was to replicate the general finding of superiority in visual search for a clinical group diagnosed with ASD. Furthermore, the aim was to establish a baseline of superior performance in visual search to be compared with intergroup effects in parallel studies. However, the results of experiment 1 indicated no such superior performance for the ASD group compared to matched controls. To uncover why this difference in performance was absent, a second experiment was conducted, which more closely resembled the Kemner et al. (2008) study. A visual search display varying in set size (i.e. 4, 16 or 25 elements) was used, and the experiment consisted of two separate sessions. In one session the target was a tilted line among vertical line distractors, referred to as the ‘easy’ session, and in the other session the target was a vertical line among tilted line distractors, referred to as the ‘hard’ session. The naming of these sessions (i.e. ‘easy’ or ‘hard’) is in accordance with earlier research using such visual search displays (Kemner et al., 2008; O’Riordan et al., 2001). All but one individual with ASD and two controls had previously participated in the first experiment.

Results from the second experiment were in accordance with general findings in the visual search literature. Reaction times increased when the target was absent, and reaction times also increased with set size, particularly when the target was absent. As in the Kemner et al. (2008) study, in the ‘easy’ session, the ASD group was significantly faster than the matched

## 6. *Limit to search superiority in ASD*

control group, especially when the target was absent. Furthermore, this difference between the ASD group and the control group when the target was absent increased as set size became larger. In the ‘hard’ session, however, the superior performance for the ASD group was not significant. Why this might be the case will be discussed later on. For now, the results from experiment 1 indicate no differences between the ASD group and the control group in visual search performance, whereas the results from experiment 2 do demonstrate superior visual search performance for the ASD group compared to the control group.

In order to put these results in context, we have to consider the differences between experiment 1 and 2. Experiment 2 was set up to match the Kemner et al. (2008) study, with two sessions (i.e. ‘easy’ and ‘hard’ session) and three set sizes in each session. The tilted lines, which could either be distractors or the target depending on the session, were tilted  $17^\circ$  clockwise from vertical. Experiment 1, however, only contained one session, in which the target was a vertical line among tilted line distractors. The tilt of the line was smaller, at only  $10^\circ$  clockwise from vertical, making it more similar to the target. Furthermore, only the set size of 25 elements was included. Finally, it is important to note that the participants in experiment 2 had all but three (1 in the ASD and 2 in the control group) participated in experiment 1, with at least six months in between measurements, although the group size was somewhat smaller. In other words, the differences between experiment 1 and 2 are limited to time of measurement, group size and differences in the visual search display (i.e. one versus three set sizes and decreased tilt of lines).

In order to discuss the results of the present experiments in terms of experimental differences (i.e. changes in target-distractor similarity and the addition of set size as a factor), first we must rule out other possible explanations for the observed effects. As the group of participants in both experiments is highly similar, one could suggest that some form of perceptual learning between sessions has taken place. However, this would suggest that perceptual learning had taken place in the ASD group only, and span

at least six-months. Research on perceptual learning in individuals with ASD shows that controls outperformed the group diagnosed with ASD on a perceptual learning task they have previously been exposed to (Plaisted, O’Riordan, & Baron-Cohen, 1998b). This would then suggest that, if perceptual learning had taken place, the control group would outperform the ASD group in the current study, which was not the case and is thus not likely to explain the current results. Another possibility is that there is something special about the clinical group in the current study. So far, two studies have previously reported no superior performance in a visual search task for individuals with ASD. First (Constable, Solomon, Gaigg, & Bowler, 2010) found no differences in a visual search task requiring participants to find an ellipse in a set of circles between the clinical group and the control group, both in reaction time and error rates. They suggest this might be due to an older age group compared to previous studies, which might indicate that perceptual differences between individuals with ASD and typically developing individuals could diminish over time (Constable et al. 2010). As the age of the participants in the present study resembles that of previous research (Kemner et al., 2008; O’Riordan, 2004), and superior performance was actually observed in experiment 2 (similar to Kemner et al., 2008), this suggestion is not applicable to the present study. Second, Baldassi et al. (2009) reported no superiority for locating a deviant stimulus at the center of vision for a group of children with ASD compared to typically developing children. However, as the task in Baldassi et al. (2009) measured sensitivity thresholds for briefly presented sets of stimuli, it is hard to compare their results to the present study. The remaining option is that the difference in results between experiments 1 and 2 is due to differences in the visual search task itself.

As previously stated, the visual search task in experiment 1 differed from that in experiment 2 on a number of factors: only one session was included (i.e. vertical target among tilted distractors), one set size (i.e. 25 elements) was presented, and the distractors were more similar to the target (tilted only 10° clockwise from vertical instead of 17°). Whereas the two sessions in experiment 2 were referred to as ‘easy’ and ‘hard’, this categorization of

## 6. *Limit to search superiority in ASD*

difficulty cannot easily be extended to experiment 1. Increasing difficulty on a specific task will undoubtedly result in slower reaction times, that is, if the error rates remain constant. However, slower reaction times (at constant error rates) do not necessarily reflect a more difficult task if the task is not identical. The fact that the number of elements in the visual search display, or set size, was not varied in experiment 1, meant that switching between different set sizes was not necessary. This, in itself, makes the task in experiment 1 different from the task in experiment 2, and makes comparison of task difficulty troublesome.

Something can, however, be said about the perceptual load of the visual search displays in both tasks, instead of task difficulty. Lavie & de Fockert (2003) provide a clear distinction of task difficulty and perceptual load, and the two should not be confused when interpreting the following explanation. As the name implies, perceptual load is the concept of how much load is being placed on the perceptual system, something that increases as both the number of elements in a visual search display is increased as well as when target-distractor similarity is increased (Lavie, 1995; Lavie & Cox, 1997). This concept of perceptual load provides the basis for Load Theory, which posits that human perception has limited capacity and automatically uses that capacity for perceptual processing (Lavie, 2005; Lavie, Hirst, de Fockert, & Viding, 2004). If perceptual load on a task is low such that the task does not require all capacity, any capacity remaining is automatically used for task-irrelevant information. When perceptual load on a task is high, no capacity is left, and task-irrelevant information is not processed. This is demonstrated by the decreased interference of distractors when perceptual load is high (Lavie, 1995; Lavie & Cox, 1997). What this means for the current experiments is that perceptual load increases when set size increases, which is the case in experiment 2. Furthermore, perceptual load also increases when target-distractor similarity is increased; thus the visual search display in experiment 1 has a higher perceptual load than the largest (i.e. 25 elements) visual search display in experiment 2.

A recent study by Remington, Swettenham, & Lavie (2012) suggests that high perceptual load impairs visual detection in typically developing adults, but not in adults diagnosed with ASD. They suggest that superior performance typically observed for individuals with ASD in visual search (Joseph et al., 2009; Kemner et al., 2008; O’Riordan et al., 2001; O’Riordan, 2004; Plaisted, O’Riordan, & Baron-Cohen, 1998a) is related to an enhanced perceptual capacity in individuals with ASD. This enhanced capacity implies that individuals with ASD process more information, and although that additional information can be task-irrelevant, it can also be task-relevant (Remington, Swettenham, & Lavie, 2012). In this case, the task-relevant information would mean the detection of a target among distractors. Previously, Remington, Swettenham, Campbell, & Coleman (2009) demonstrated that individuals with ASD require higher levels of perceptual load to be able to ignore irrelevant distractors, again suggesting enhanced capacity for individuals with ASD. Furthermore, this enhanced capacity seems to be reflected in the severity of autistic traits, as typically developing individuals scoring high on the AQ are affected less by higher perceptual load than individuals scoring low on the AQ (Bayliss & Kritikos, 2011). While this concept of enhanced perceptual capacity has been used to explain superior performance in individuals with ASD, it might also explain the results observed in experiment 1 of the present study. A possible explanation for the results in the present study is that the perceptual load of the display in experiment 1 has become so high that also the capacity of the individuals with ASD has left no room for the processing of additional task-(ir)relevant information. As such, the critical factor is no longer in capacity for processing multiple information sources, but in the discrimination of a single stimulus-element. With reference to the search displays in the present study this would be mean that the task in experiment 1 is so demanding that both participants with ASD and controls make a decision of target or distractor for one element at a time. In cases of lower perceptual load, as in experiment 2, multiple stimulus-elements can be processed at the same time, with the enhanced perceptual capacity of individuals with ASD allowing for more stimulus-elements to be processed simultaneously.

## 6. *Limit to search superiority in ASD*

If perceptual capacity for both the ASD and control group leaves no room for processing more than one stimulus-element at a time, any superiority for either group would be due to static stimulus discrimination or identification. Previous research has demonstrated that there is no clear evidence for superior or inferior performance in static stimulus discrimination in ASD, which might contribute to the result in the present study (Simmons et al., 2009). For instance, De Jonge et al. (2007) found no differences in contrast sensitivity and thresholds for form discrimination between individuals with ASD and controls. Although Bertone, Mottron, Jelenic, & Faubert (2005) demonstrated lower thresholds for orientation detection (i.e. detecting whether a stimulus is oriented one way or another) in individuals with ASD under certain circumstances, no differences have been reported on orientation discrimination (i.e. discriminating the tilt of a line from another) between individuals with ASD and controls. This suggests that when the task breaks down to single stimulus-element discrimination, no differences are to be expected, and could explain the results of experiment 1 in the present study.

With reference to the previously noted study of Constable et al. (2010), this could mean that the absence of a difference between individuals with ASD and matched-controls in their visual search task might not be due to an older age group, but to limits on perceptual load and the consequent stimulus discrimination. Whether perceptual load in the task Constable et al. (2010) use (i.e. finding an ellipse in a set of distractor circles) is actually higher than when finding a tilted line among vertical line distractors has thus far not been tested directly. Future research should prove useful in determining perceptual load across different paradigms. It should be noted, however, that the explanation Constable et al. (2010) provide for their results might still very well be accurate; that perceptual differences between individuals with ASD and controls can diminish with age.

When the findings of the present study are put in the context of load theory we can hypothesize the following; as long as perceptual capacity leaves room for processing of additional task-(ir)relevant information, indi-

viduals with ASD will show superior performance. When perceptual load is so high that no room is left for processing additional task-(ir)relevant information, even in individuals with ASD, no differences will be found on static stimuli between individuals with ASD and controls. Future research will be helpful in testing these claims, pinpointing the scope and limits of the superiority in visual search for individuals with ASD, and uncovering the specific anomalies of visual processing in ASD.

A final note on the results of experiment 2 in the present study is in order. Although the results from the ‘easy’ session in experiment 2 showed superiority in reaction times for the group diagnosed with ASD compared to controls, this effect was not found in the ‘hard’ session. Numerically, the effect does appear to be present, even larger than in the ‘easy’ session, though the effects are not significant. A possible explanation for the lack of statistical significance is the large standard errors in the control group in the ‘hard’ condition. When we compare the present results with the Kemner et al. (2008) study, whose experiment most closely resembles experiment 2 in the present study, similar standard errors are observed (i.e. around 200 ms in the larger set sizes when the target is absent). Mean reaction times for the ASD group in the fastest conditions (i.e. ‘easy’ session when the target is present) are around 600 ms in both studies, although mean reaction times for the control group in the slowest conditions vary; around 1350 ms in the present study compared to 1600 ms in the Kemner et al. (2008) study. The absence of a significant difference might then be due to a more condensed spread in the reaction times, though the standard errors are quite similar. Furthermore, as the AQ scores in the control population are quite high, this might have lessened the difference between the ASD group and the control group (see e.g. Bayliss & Kritikos, 2011). Nevertheless, these considerations do not undermine the finding of superiority in visual search for individuals diagnosed with ASD, as reflected in the ‘easy’ session. As such, the absence of statistical significance in the ‘hard’ session does not deter from the conclusions of the present study and the hypothesis regarding perceptual load as a limiting factor for visual search superiority in ASD.

## **Acknowledgements**

We thank Branka Milivojevic, Emmie van Schaffelaar, Manje Brinkhuis, Carlijn van den Boomen, and Esther Eijlers for their help with participant recruitment and data collection, and Siarhei Uzunbajakau for building the experimental setup. This work was supported by a Netherlands Organization for Scientific Research (NWO) VICI grant (45307004) to Chantal Kemner.

## References

- Almeida, R. A., Dickinson, J. E., Maybery, M. T., Badcock, J. C., & Badcock, D. R. (2010). A new step towards understanding embedded figures test performance in the autism spectrum: The radial frequency search task. *Neuropsychologia*, 48(2):374–381.
- Bach, M. (1996). The Freiburg visual acuity test—automatic measurement of visual acuity. *Optometry and Vision Science*, 73(1):49–53.
- Baldassi, S., Pei, F., Megna, N., Recupero, G., Viespoli, M., Igliozzi, R., Tancredi, R., Muratori, F., & Cioni, G. (2009). Search superiority in autism within, but not outside the crowding regime. *Vision Research*, 49(16):2151–2156.
- Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The autism-spectrum quotient (AQ): Evidence from asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders*, 31(1):5–17.
- Bayliss, A. P., & Kritikos, A. (2011). Brief report: Perceptual load and the autism spectrum in typically developed individuals. *Journal of Autism and Developmental Disorders*, 41(11):1573–1578.
- Bertone, A., Mottron, L., Jelenic, P., & Faubert, J. (2005). Enhanced and diminished visuo-spatial information processing in autism depends on stimulus complexity. *Brain*, 128:2430–2441.
- Constable, P. A., Solomon, J. A., Gaigg, S. B., & Bowler, D. M. (2010). Crowding and visual search in high functioning adults with autism spectrum disorder. *Clinical Optometry*, 2:93–103.
- Dakin, S., & Frith, U. (2005). Vagaries of visual perception in autism. *Neuron*, 48(3):497–507.
- De Jonge, M. V., Kemner, C., de Haan, E. H., Coppens, J. E., van den Berg, T. J. T. P., & van Engeland, H. (2007). Visual information processing in high-functioning individuals with autism spectrum disorders and their parents. *Neuropsychology*, 21(1):65–73.
- Elsabbagh, M., Volein, A., Csibra, G., Holmboe, K., Garwood, H., Tucker, L., Krljes, S., Baron-Cohen, S., Bolton, P., Charman, T., Baird, G., & Johnson, M. H. (2009). Neural correlates of eye gaze processing in the infant broader autism phenotype. *Biological Psychiatry*, 65(1):31–38.
- Hooge, I. T. C., & Erkelens, C. J. (1999). Peripheral vision and oculomotor control during visual search. *Vision Research*, 39:1567–1575.
- Jarrold, C., Gilchrist, I. D., & Bender, A. (2005). Embedded figures detection in autism and typical development: Preliminary evidence of a double dissociation

## 6. Limit to search superiority in ASD

- in relationships with visual search. *Developmental Science*, 8(4):344–351.
- Joseph, R. M., Keehn, B., Connolly, C., Wolfe, J. M., & Horowitz, T. S. (2009). Why is visual search superior in autism spectrum disorder? *Developmental Science*, 12(6):1083–1096.
- Kaldy, Z., Kraper, C., Carter, A. S., & Blaser, E. (2011). Toddlers with autism spectrum disorder are more successful at visual search than typically developing toddlers. *Developmental Science*, 14(5):980–988.
- Kemner, C., Ewijk, L., Engeland, H., & Hooge, I. (2008). Brief report: Eye movements during visual search tasks indicate enhanced stimulus discriminability in subjects with PDD. *Journal of Autism and Developmental Disorders*, 38(3):553–557.
- Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception and Performance*, 21(3):451–468.
- Lavie, N. (2005). Distracted and confused?: Selective attention under load. *Trends in Cognitive Sciences*, 9(2):75–82.
- Lavie, N., & Cox, S. (1997). On the efficiency of visual selective attention: Efficient visual search leads to inefficient distractor rejection. *Psychological Science*, 8(5):395–398.
- Lavie, N., & de Fockert, J. W. (2003). Contrasting effects of sensory limits and capacity limits in visual selective attention. *Perception & Psychophysics*, 65(2):202–212.
- Lavie, N., Hirst, A., de Fockert, J. W., & Viding, E. (2004). Load theory of selective attention and cognitive control. *Journal of Experimental Psychology: General*, 133(3):339–354.
- Lord, C., Rutter, M., & Couteur, A. (1994). Autism diagnostic interview-revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of autism and developmental disorders*, 24(5):659–685.
- Lord, C., Rutter, M., Goode, S., Heemsbergen, J., Jordan, H., Mawhood, L., & Schopler, E. (1989). Autism diagnostic observation schedule: A standardized observation of communicative and social behavior. *Journal of Autism and Developmental Disorders*, 19(2):185–212.
- Merin, N., Young, G. S., Ozonoff, S., & Rogers, S. J. (2007). Visual fixation patterns during reciprocal social interaction distinguish a subgroup of 6-month-old infants at-risk for autism from comparison infants. *Journal of Autism and Developmental Disorders*, 37(1):108–121.
- O’Riordan, M. A. (2004). Superior visual search in adults with autism. *Autism*,

8(3):229–248.

- O’Riordan, M. A., Plaisted, K. C., Driver, J., & Baron-Cohen, S. (2001). Superior visual search in autism. *Journal of Experimental Psychology: Human Perception and Performance*, 27(3):719–730.
- Pellicano, E., Smith, A. D., Cristino, F., Hood, B. M., Briscoe, J., & Gilchrist, I. D. (2011). Children with autism are neither systematic nor optimal foragers. *Proceedings of the National Academy of Sciences*, 108(1):421–426.
- Plaisted, K., O’Riordan, M., & Baron-Cohen, S. (1998a). Enhanced visual search for a conjunctive target in autism: A research note. *Journal of Child Psychology and Psychiatry*, 39(5):777–783.
- Plaisted, K., O’Riordan, M., & Baron-Cohen, S. (1998b). Enhanced discrimination of novel, highly similar stimuli by adults with autism during a perceptual learning task. *Journal of Child Psychology and Psychiatry*, 39(5):765–775.
- Remington, A., Swettenham, J., Campbell, R., & Coleman, M. (2009). Selective attention and perceptual load in autism spectrum disorder. *Psychological Science*, 20(11):1388–1393.
- Remington, A. M., Swettenham, J. G., & Lavie, N. (2012). Lightening the load: Perceptual load impairs visual detection in typical adults but not in autism. *Journal of Abnormal Psychology*, 121(2):544–551.
- Simmons, D. R., Robertson, A. E., McKay, L. S., Toal, E., McAleer, P., & Pollick, F. E. (2009). Vision in autism spectrum disorders. *Vision Research*, 49(22):2705–2739.



## 7. An in-depth look at saccadic search in infancy

Published as:

Hessels, R. S., Hooge, I. T. C., & Kemner, C. (2016). An in-depth look at saccadic search in infancy. *Journal of Vision*, 16(8):10.

Author contributions:

RH, IH, CK designed the study. Data was collected by RH and research assistants at the KKC under supervision of RH. RH, IH analyzed the data. RH, IH, CK interpreted the data. RH drafted the paper. RH, IH, CK finalized the paper.

## **Abstract**

Two questions were posed in the present study: 1) Do infants search for discrepant items in the absence of instructions? We outline where previous research has been inconclusive in answering this question. 2) In what manner do infants search, and what are the fixation and saccade characteristics in saccadic search? A thorough characterization of saccadic search in infancy is of great importance as a reference for future eye-movement studies in infancy. We presented 10-month-old infants with 24 visual search displays in two separate sessions within two weeks. We report that infant saccadic search performance at 10 months is above what may be expected by our model of chance, and is dependent on the specific target. Infant fixation and saccade characteristics show similarities to adult fixation and saccade characteristics in saccadic search. All findings were highly consistent across two separate sessions on the group level. An examination of the reliability of saccadic search revealed that test-retest reliability for oculomotor characteristics was high, particularly for fixation duration. We suggest that future research into saccadic search in infancy adopt the presented model of chance as a baseline against which to compare search performance. Researchers investigating both the typical and atypical development of visual search may benefit from the presented results.

For humans and other primates, it is vital to adequately process their visual surroundings. One method that has been prevalent in studying (aspects of) processing of a visual scene in both humans and primates is studying behavior in a visual search task (e.g. Motter & Belky, 1998a; Williams, 1967). The basics of a visual search task are straightforward; subjects are required to search for a target stimulus in a visual scene, which may or may not be present (Wolfe, 1998a). The subject then commonly responds whether the target is present or not. The dependent variables are often accuracy and reaction time of target detection. The target is defined by the presence or absence of a feature or some unique combination of features, and is located somewhere between non-target stimuli. Whilst there are many variants of the visual search task, the overarching concept is that by varying aspects of both the target and non-target one might investigate different aspects of visual processing (Wolfe, 1998a). For example, visual search tasks have been used extensively to study and model attentional processes (Treisman & Gelade, 1980; Wolfe, 1994) and oculomotor control (Hooge & Erkelens, 1996; McPeck, Skavenski, & Nakayama, 2000; Motter & Belky, 1998b). As a visual search task can be used to study different aspects of visual processing, it may be an ideal candidate to study the development of vision from infancy to adulthood. Indeed, recent years have seen an increase in developmental studies using a visual search paradigm to study the early development of vision. We first briefly introduce an example of how research on visual search in infancy may inform developmental models of vision, after which we introduce our scope on infant eye-movement behavior in search.

Previous research into the development of processes underlying visual search, particularly in infancy, has focused on establishing whether findings reported in the adult literature are present in infancy as well. One example is visual pop-out: when the speed of target localization in visual search is relatively unaffected by the amount of non-targets. This is also referred to as pre-attentive search or efficient search (Duncan & Humphreys, 1989; Wolfe, 1998b). A specific example is searching for a feature-present target among feature-absent non-targets; for instance a Q among O's (the

## 7. *Saccadic search in infancy*

added line of the Q being the feature; Wolfe, 1998a). As the feature ‘pops out’, search time is relatively unaffected by the amount of feature-absent non-targets. However, searching for a feature-absent target among feature-present non-targets (O among Q’s) is typically inefficient: the absence of a feature does not “pop out”. This is referred to as a search asymmetry – the situation where target A among non-targets B is detected faster than target B among non-targets A. Early research in infancy first investigated whether infants detected this feature-present stimulus among feature-absent stimuli at all. Colombo, Ryther, Frick, & Gifford (1995); Quinn & Bhatt (1998) reported that infants aged 3 to 4 months preferentially looked at feature-present among feature-absent stimuli (Q among O’s), after being familiarized with just feature-absent stimuli (only O’s), but not vice versa (O among Q’s). Colombo et al. (1995); Quinn & Bhatt (1998) claimed that visual pop-out was present in infancy, based on this asymmetry in preferentially looking at feature-present but not feature-absent stimuli. However, Adler & Orprecio (2006) argued that this could not be concluded on the basis of preferential looking times, which are typically on the order of several seconds, while visual pop-out or efficient search occurs on the order of several hundreds of milliseconds. They argued that knowledge of the eye movements was needed in order to draw this conclusion. In their study, Adler & Orprecio (2006) measured saccade latencies of 3 month-old infants to a target and report that saccade latency was unaffected by the number of non-targets. When search for a target is unaffected by the number of non-targets, it may be noted as visual pop-out or efficient (Wolfe, 1998a). Adler & Orprecio (2006) consequently conclude that visual pop-out is present in infancy. Adler & Gallego (2014) extend this by reporting that infants exhibit efficient search for feature-present but not for feature-absent visual search trials. Concluding, there is evidence that infants exhibit visual pop-out and search asymmetries, and these findings may consequently inform developmental models of visual search (e.g. by extrapolating brain-based models to infants; Li, 1999), or more general models of visual attention (e.g. M. H. Johnson, 1990).

In the present study we take an in-depth look at eye-movement behavior itself – specifically in a highly relevant visual task, namely search. We are concerned with saccadic search in infancy, and how saccadic search behavior in infancy can be characterized. Saccadic search is the topic in a subset of the visual search literature, and is concerned with target localization after multiple saccades (Caspi, Beutter, & Eckstein, 2004; Hooge, Over, van Wezel, & Frens, 2005; Motter & Belky, 1998a; 1998b; Scinto, Pillalamarri, & Karsh, 1986; Vlaskamp & Hooge, 2006; Vlaskamp, Over, & Hooge, 2005; Wu & Kowler, 2013). Saccadic search may occur, for example, when instructions are given to find a target, but the target cannot be located at first glance. Here, search is an active exploration of the visual scene as opposed to mere visual pop-out (i.e. locating the target at first glance). The investigation of saccadic search in infancy is particularly relevant for multiple reasons. First, by investigating eye-movements in the context of search, one can draw conclusions about performance (i.e. how well do infants localize the target?). For infants, this is particularly interesting due to the absence of instructions: do infants actually search for a discrepant item in the absence of instructions? If infants do, saccadic search may be used to investigate the development of oculomotor control. Second, a characterization of eye-movement behavior in search in infancy may allow us to compare infant saccadic search to primate and adult saccadic search in the future. This comparison may, for instance, shed light on whether, and how, search strategies are learned or innate, and how they develop. Moreover, a thorough characterization may serve as a reference for distinguishing typical from atypical development of visual search behavior, as has previously been observed in Autism Spectrum Disorder (ASD), for example Gliga et al. (2015); Kaldy, Kraper, Carter, & Blaser (2011); see also the suggestions made in Hessels, Hooge, Snijders, & Kemner (2014). For a recent review on visual search in ASD see Kaldy, Giserman, Carter, & Blaser (2016).

Investigating saccadic search in infancy is, however, not as straightforward as in adults. While adults can be given instructions to find a target, infants cannot (or will not follow instructions when given so). One approach adopted by Kaldy et al. (2011) in older children (toddlers of around 2.5

## 7. *Saccadic search in infancy*

years old) is to indicate the special status of the target by familiarizing children with it before search trials commenced. Another approach is to present infants with stimuli that are typically used to study search behavior in adults, and determine whether infants spontaneously search for a discrepant item. As no instructions are given, we deem it more appropriate to refer to visual search stimuli instead of a visual search task when studying search in infancy. One major assumption here is that, when infants fixate the target, infants have identified it as being a discrepant item. Indeed, for toddlers there is sound evidence that the target was indeed perceived to be special in the absence of instructions (Kaldy et al., 2011). In recent years, two studies have used visual search stimuli in infancy that may have elicited saccadic search, while eye movements were simultaneously recorded. Amso & Johnson (2006) presented 3-month-old infants with search displays containing vertical lines and one titled line (the target). In addition, they presented search displays containing vertical lines and one moving line (the target) as a visual pop-out control condition. They report that infants fixated the static target in about half the trials compared to nearly 90% of the moving targets. The average time to target hit (the first time the target was fixated) for the static target was around 1400-1500 ms, compared to roughly 1100 ms for the moving target. Frank, Amso, & Johnson (2014) employed the same visual search displays for 3, 6, and 9-month-old infants, and report that the proportion of target hits – defined as fixation of the target location, measured with an eye tracker – in the static trials increases from roughly 25 to 40% from 3 to 9 months. For the trials containing moving targets, the proportion of target hits increases from roughly 55 to 100%. Moreover, time to target hit for the moving trials decreased from roughly 1800 to below 1000 ms, whereas time to target hit for the static trials did not decrease as a function of age. These two studies clearly demonstrate that target localization is slower for static targets compared to moving targets. Moreover, Amso & Johnson (2006) conclude that target localization is better than what may be expected by chance.

While Amso & Johnson (2006); Frank et al. (2014) conclude that infants indeed search for a target, even in the absence of instructions, there are two

questions that need to be further addressed. First, given that there were 8 possible target locations, Amso & Johnson (2006) expected infants to locate the target by chance in 12.5% of the cases. This model might be valid when there are 8 possible options, and one must be chosen. However, infants were given a maximum of 4 seconds to locate the target. It may very well be that multiple possible target-locations were selected in that four-second period. Moreover, it may also be that no target-locations were selected at all. It may therefore not be the case that one, and only one, of 8 possible locations is fixated. We consequently deem the model of 12.5% chance unfit for this experiment, and propose a different model of chance. The model we propose is derived from the definition of saccadic search: target localization that generally occurs after multiple saccades. Saccadic search then consists of two parts: explorative behavior – fixations on different parts of the display – and orienting to the target – in other words, fixating the target. During the first part (explorative behavior), any number of non-targets in the display may be fixated. If the target has no special status over the non-targets, one would not expect that the target is fixated with higher probability than non-targets in the same time span. However, given that there is a general bias to fixate in the center of the screen, the probability of each non-target being fixated is not necessarily equal, and consequently it is impossible to determine the probability of fixating any particular element in a display up front. However, by computing the search time and probability of fixating non-targets at equal distance from the center of the screen as the target, we obtain a chance level. For example, if a non-target at the same distance from the center as the target is localized as often in the same time span as the target, target localization is not above chance. On the other hand, if the target is consistently localized faster and more often than non-targets at equal distance from the center of the screen, target localization is above chance.

The second question that remains from Amso & Johnson (2006); Frank et al. (2014), is whether time to target localization (as determined by the time from trial onset to target fixation) depends on the dissimilarity between targets and non-targets. When a target cannot be localized at first

## 7. *Saccadic search in infancy*

glance and a saccade (or multiple) is required, there are multiple aspects that may determine search performance. Increasing the number of non-target stimuli to a search display should decrease search performance (see e.g. Wolfe, 1998a). Previous work on visual search in toddlers has used this approach effectively to conclude that toddlers with ASD are more effective in visual search than typically developing controls (Kaldy et al., 2011). Moreover, if dissimilarity between target and non-targets decreases, search performance should decrease also (Duncan & Humphreys, 1989). In the previous experiments, one would therefore expect search performance to differ for the three different targets (30°, 60°, and 90°). While this is not essential to conclude that infants search, it may be expected on the basis of the adult visual search literature. These two remaining questions are addressed in the first question of the present study, and its components:

Q1: Do infants search for a discrepant item in the absence of instructions?

Q1.1 (Necessary): Is target localization above what may be expected by our model of chance?

Q1.2 (Additional): Is time to target localization dependent on target and non-target dissimilarity?

In the present study, we aim to answer this question with an experiment in a group of 10-month-old infants. Given that Amso & Johnson (2006); Frank et al. (2014) concluded above-chance target localization for infants between 3 and 9 months old, we hypothesize that infants at 10 months should also search for a discrepant item in the absence of instructions.

The second question addressed in the present study concerns the characterization of saccadic search in infancy. Previous infant studies in which saccadic search may have taken place have hitherto not reported eye-movement measures other than saccadic reaction time, and it therefore remains unknown exactly how infants search. A thorough description of infant oculomotor characteristics may serve as a reference for future eye-movement studies with infants. In addition, a characterization of infant saccadic search may be of importance to research on atypical development such as

in ASD, where visual search superiority is often reported (e.g. Kaldy et al., 2016; O’Riordan, Plaisted, Driver, & Baron-Cohen, 2001). We determine and describe oculomotor characteristics of infant saccadic behavior in visual search displays. First, fixation duration is often used as an estimate for the average visual processing time at the location fixated (Hooge & Erkelens, 1996). The processing time may subsequently be seen as an estimate for the difficulty of the visual stimulus. As in adult visual search, we expect the initial fixation duration – or the latency to initiate the first saccade – to be longer than subsequent fixation durations (Hooge & Erkelens, 1996; Zingale & Kowler, 1987). Second, we describe infant saccadic directional and amplitude changes during search as has previously been done in adults (Hooge et al., 2005).

Q2: In what manner do infants search? What are the fixation and saccade characteristics of infant saccadic search?

Given the recent emphasis on replication of psychological research (e.g. Open Science Collaboration, 2015) we presented the group of 10-month-old infants with the same visual search stimuli twice within two weeks to ascertain that the observed saccadic search behavior was reliable at the group level.

## **7.1. Method**

### **7.1.1. Participants**

Seventy-seven infants were invited into the lab center for a day of multiple studies (see e.g. Hessels, Andersson, Hooge, Nyström, & Kemner, 2015), recruited through the local municipality. Of the 77 infants invited, sufficient data (at least half the trials, mean 21.7 trials) were recorded for 55 (25 male, 30 female) infants in the first session of the visual search study. Forty (18 male, 22 female) out of the 55 that provided enough data for the first session completed the visual search study on their second visit. They all provided data for at least one third of the trials (mean 19.6 trials) on their

## 7. Saccadic search in infancy

second visit. The mean age during the first session was 302 days ( $sd = 11.8$  days); the mean age during the second session was 307 days ( $sd = 11.4$  days). Infants were only invited to participate if the parents indicated that the infants were not born preterm (i.e. before 37 weeks of pregnancy), and indicated that the infants had no impaired hearing or vision, or any developmental disorders. Parents gave written informed consent on the day of the first session, and the study was approved by the ethics committee of the local University Medical Centre (Protocol ID 14-221) and conducted in accordance with the Declaration of Helsinki. Parents received a 10 € compensation for each testing day, with another 5 € travel compensation if required.

### 7.1.2. Apparatus & Stimuli

Stimulus presentation was handled by MATLAB R2013a and the Psych-Toolbox (version 3.0.11; Brainard, 1997) running on a MacBook Pro with OS X 10.9. Stimuli were presented on an external 23-inch screen belonging to the Tobii eye tracker at a resolution of 1920 by 1080 pixels and a refresh rate of 60 Hz. The Tobii TX300 eye tracker running at 300 Hz was used for tracking infants' eye movements. The median accuracy in session 1 was  $0.89^\circ$ , and  $1.03^\circ$  in session 2. The standard deviation for accuracy values is non-informative as infants may not attend the validation stimulus and consequently achieve very large values for accuracy (see Hessels et al., 2015 for an elaborate discussion). The median precision during detected fixations (see section *Data reduction*) in session 1 was  $0.61^\circ$  RMS ( $sd = 0.36^\circ$ ), and  $0.62^\circ$  RMS ( $sd = 0.36^\circ$ ) in session 2. Further measures on data quality from the present experiment have been reported on extensively in Hessels et al. (2015). The Tobii SDK was used for communication between MATLAB and the eye-tracker.

The experiment consisted of 24 visual search displays (based on Amso & Johnson, 2006). Each visual search display consisted of 28 white lines ( $3.3^\circ$  by  $0.9^\circ$ ) as target candidates on a black background (see Figure 7.1 for a schematic version of a typical search display). The lines were arranged in a

grid of 14 columns by 2 rows, and subsequently jittered between  $-1.6^\circ$  and  $1.6^\circ$  in the horizontal and between  $-6.3^\circ$  and  $6.3^\circ$  in the vertical direction. All lines except the target line were aligned vertically. The target line was tilted  $30^\circ$ ,  $60^\circ$ , or  $90^\circ$  clockwise, and could appear in one of eight fixed locations. Each combination of target line angle and location was presented once, resulting in 24 trials. Preceding the visual search experiment was a 5-point calibration sequence. Each calibration point consisted of a colored spiral (red, green, yellow, purple, or blue) on a black background. The spiral changed in size between  $4.0^\circ$  and  $5.4^\circ$  at 0.8 Hz following a sinusoidal wave. In addition, the spiral rotated at 0.8 Hz. Following a key press of the operator, the spiral shrank in size to  $0.5^\circ$  over a period of 0.5 s. The spiral then remained on screen for 0.2 s. At the start of this period the point was calibrated. Following the first and every additional 5th visual search display a validation stimulus identical to the calibration spiral was presented to determine accuracy.

### **7.1.3. Procedure**

#### **Positioning**

The infants and parents were welcomed into the eye-tracking room and familiarized with the experimental setup. Thereafter, the infants were strapped in a baby seat, and the parent was seated on a height-adjustable chair. The baby seat was subsequently placed on the parent lap, with the infant placed parallel to the screen of the eye tracker. Positioning the infant in a baby seat was done as this would give the most stable positioning through the recording and limit the infants' movements. If, however, the parent indicated that the baby seat would probably result in a restless or upset infant, the infant was seated without a baby seat in the parents lap or in a high chair. The decision for either the parents lap or the high chair was up to the judgment of the operator, i.e., which of the two would work best for the particular infant. After positioning the parent and infant, the position of the eye tracker was adjusted so that the eyes of the infant were at 65 cm from the eye tracker and at the same height as the center of the screen.

## 7. Saccadic search in infancy

### Calibration and experiment

After positioning, a 5-point calibration sequence was started. Calibration stimuli were serially presented in the four corners and center of the screen. The order of points was random each time the calibration was run. The infant was monitored with a webcam. The operator judged from this video whether the infant looked in the direction of the calibration stimulus and pressed the spacebar to calibrate the current point. After the calibration sequence the calibration output was examined. Calibration points without data, or with data that were inconsistent and characterized by dispersed gaze points around the calibration point were re-calibrated by the operator. Each re-calibration was noted down as an additional calibration run. After calibration was deemed successful, or when the infant started losing attention, the experiment was initiated. A colorful static picture was presented centrally before each visual search trial to attract or maintain the infants' attention. The operator initiated the trial by pressing the spacebar when the infant was judged to look at the screen. The visual search trial remained on screen until either 4 seconds had passed, or the infant fixated the target within a range of  $1.4^\circ$  for a minimum duration of 100 ms. This was very conservative, and only allowed precise target fixations to end the trial. Note that the definition for target hit is different in the offline analysis (see section *Data reduction*). After the trial ended, a short video clip of a popular cartoon was presented at the target location in an attempt to stimulate spontaneous visual search. If, during the experiment, the infant was not attending to the screen, the operator could present sounds, or videos (accompanied with sound) in the center of the screen, to attract the infants' attention. The entire experiment, including calibration and positioning, lasted approximately 10 to 15 minutes.

#### 7.1.4. Data reduction

Raw position signals from the left and right eye were first combined into an average position signal. If gaze position was only available from one eye, that signal was used. Cubic spline interpolation was performed for periods of data loss with a duration of less than 100 ms (Frank, Vul, & Johnson,

2009). Hereafter, a fixation detection algorithm specifically designed for use in infant data was applied. The algorithm operates as an adaptive dispersion algorithm, with which fixation detection can be achieved across large variations in noise levels, both local and between participants or trials. The algorithm, Identification by 2-Means Clustering (I2MC; Hessels, Niehorster, Kemner, & Hooge, 2016), is based on a procedure called *k*-means clustering (where  $k = 2$ ), which is used to determine whether one or two fixation clusters are present in a small moving window (200 ms in this case). When the window contains a saccade, the algorithm detects two clusters in 2D space (fixations) that are clearly separated in time (i.e. one cluster follows the next in time). When the window contains no saccade, the algorithm is still forced to detect two clusters in 2D space, although they are not clearly separated in time. The clustering method combined with a moving window ensures a noise-robust detection of fixations, both within and between participants. Finally, fixations shorter than 40 ms were excluded, and successive fixations separated less than  $0.7^\circ$  in location and less than 30 ms in time were merged.

In infant research, the amount of data available per infant may vary widely due to e.g. inattention by some infants, or differences in eye-tracking stability (Hessels et al., 2015). As we were mainly interested in determining saccadic search behavior of 10-month-old infants in general, trials from all infants were pooled following fixation detection in order to produce better estimates for group search performance measures (see e.g. Adler & Gallego, 2014; Adler & Orprecio, 2006). Hereafter, time to target hit, number of fixations to target hit, and proportion of trials in which the target was hit were calculated. Target hit was defined as gaze entering an area of  $4.7^\circ$  from the center of the target.

In order to determine a baseline against which search performance for the target could be compared, we ran a computation for comparison non-targets. In each visual search trial, the target could appear at one of eight possible locations: two possible locations in each quadrant of the screen. In each of the three quadrants not containing the target, a non-target was

## 7. Saccadic search in infancy

selected as a comparison non-target. For each quadrant, the non-target with the distance to the center that most closely matched the distance of the target to the center was selected. For each of the three comparison non-targets in a visual search trial, the time it took the infant to fixate that element was determined. If the number of times a target was hit in the 4-second span of a trial did not exceed the number of times a comparison non-target was hit in the same time span, we conclude that target localization is not above chance.

Although there were three comparison non-targets available for each visual search trial, there was no difference in the number of times these targets were hit in a 4-second time span: the empirical cumulative proportion of object hit as a function of time overlapped for the three non-targets. Therefore, the most conservative comparison non-target was chosen as the reference: the comparison non-target in the opposite quadrant from the target in the 90° target trials was chosen. In addition, comparison non-targets were averaged for the analyses of proportion of target-directed saccades and target capture at next saccade (see section *Saccadic search performance*).

## 7.2. Results

### 7.2.1. Saccadic search performance

Figure 7.1 depicts an example trial of one infant. As can be seen, gaze started in the center of the display and after some time, following fixation of two non-targets, the target was located in the bottom left corner. As can be seen from the left panel in Figure 7.2, the proportion of trials with a target hit was lowest for the 90° target, followed by the 30°, and finally the 60° target. In addition, the mean number of fixations to target hit (middle panel Figure 7.2) was highest for the 90° target, followed by the 30°, and finally the 60° target. Finally, the mean time to target hit was longest for the 90° target, again followed by the 30° target, and the 60° target. This relative pattern was identical across session 1 and 2, although targets were hit faster and more often in session 2. As standard statistical tests cannot

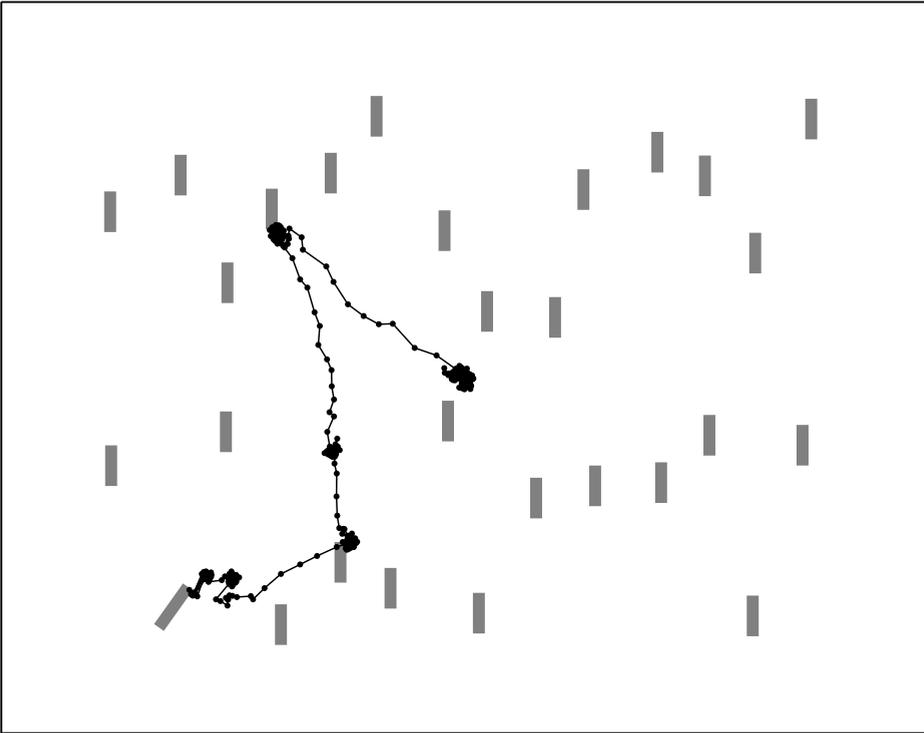


Figure 7.1.: Example visual search display with gaze overlaid in black. Gaze starts in the center and ends at the target location. The display is schematic; size and color differed in the experiment (see section *Apparatus & stimuli*).

## 7. Saccadic search in infancy

combine both proportion of target hit and time to target hit as a single measure of performance, we performed bootstrap analyses on cumulative frequencies of target hit as a function of time (see e.g. Hooge & Camps, 2013). To determine the proportion of objects that were hit after each 500 ms increment, and corresponding 95% confidence intervals, 1000 samples of equal size were drawn with replacement from the distribution of time to first hit. The non-target in the opposite quadrant in the 90°-target trial served as the chance level (see section *Data reduction*).

The cumulative proportion of target and comparison non-target hit as a function of time are depicted in Figure 7.3. As can be seen from the left panel in Figure 7.3, 60° targets were detected most often in the 4-second timeframe, followed by the 30° target, in session 1. The cumulative proportion of 90° target hit almost overlaps with the cumulative proportion of the comparison non-target hit. This suggests that the 90° target was not distinguished from non-targets above chance – as determined by the time to fixate a comparison non-target at the opposite location. For session 2, 60° targets were again detected most often in the 4-second timeframe, followed by the 30° target. This time, the cumulative proportion of 90° target hit did not overlap with the cumulative proportion of comparison non-target hit. This suggests that in session 2, the 90° target was, in fact, more often fixated than a comparison non-target.

We reasoned that if infants search actively for the target, the proportion of saccades that go towards the target might initially be low if the target is not yet locatable from the periphery. Subsequently, the proportion of saccades towards the target might increase after gaze is brought closer towards the target. We calculated the proportion of target-directed saccades as a function of saccade number to investigate this. Target-directed saccades were defined as saccades with an unsigned angle of less than 45° between the line through the saccade start and endpoint, and the line through the saccade start point and the target. Figure 7.4 depicts the proportion of target-directed saccades for the 30°, 60°, 90°-target and the mean of the comparison non-targets. As can be seen from the left panel in Figure 7.4,

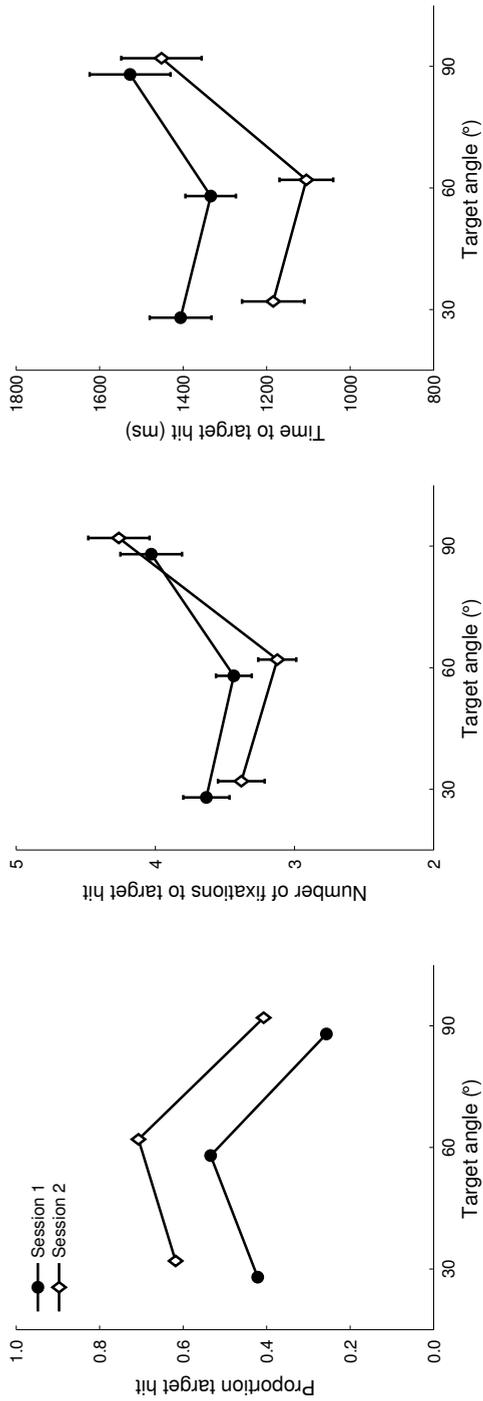


Figure 7.2.: Proportion target hit (left), number of fixations to target hit (middle), and time to target hit (right) as a function of target angle for both sessions. Error bars (middle and right panel) depict standard error of the mean. The data between the two sessions are shifted horizontally to prevent overlapping error bars.

## 7. Saccadic search in infancy

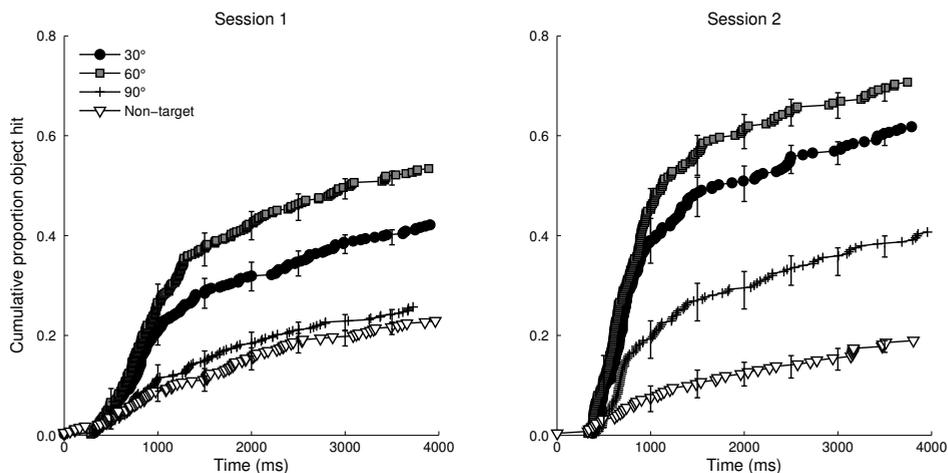


Figure 7.3.: Cumulative proportion of 30°, 60°, and 90° targets hit and comparison non-target hit as a function of time for session 1 (left) and session 2 (right). Error bars represent 95% confidence intervals acquired using bootstrapping procedures in MATLAB.

the proportion of target-directed saccades in session 1 increased after the first saccade for the 30° and 60° target, and after the second saccade for the 90° target, compared to the comparison non-target. As visible from the right panel in Figure 7.4, the proportion of target-directed saccades increased for the 30°, 60°, and 90° after the first saccade in session 2, compared to the comparison non-target. Unlike the data from session 1, the proportion of target-directed saccades was already much higher for the 30° and 60° target at the first saccade compared to the comparison non-target. These findings suggest that infant actively searched for a target beyond the first shift of gaze.

While search performance depended on the angle of the target (i.e. 30°, 60° or 90°), 90° targets were less often fixated than the other two targets. If target to non-target similarity decreases linearly with increasing target angle, a linear decrease of time to target hit as a function of target angle would similarly be expected. As this was not the case in both sessions, we

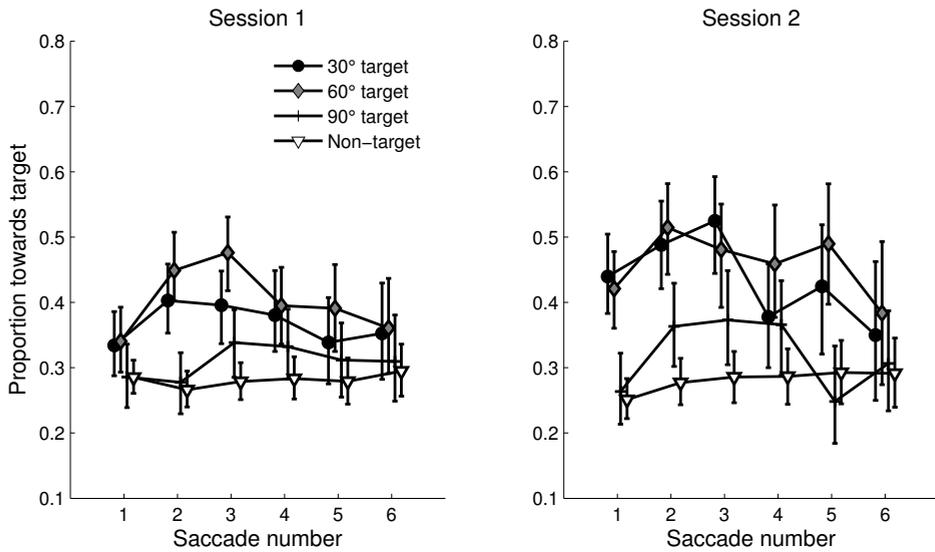


Figure 7.4.: Proportion 30° target, 60° target, 90° target, and comparison nontarget-directed saccades as a function of saccade number for session 1 (left) and session 2 (right). Error bars represent 95% confidence intervals acquired using bootstrapping procedures in MATLAB. The data between the (non-)target(s) are shifted horizontally to prevent overlapping error bars.

## 7. Saccadic search in infancy

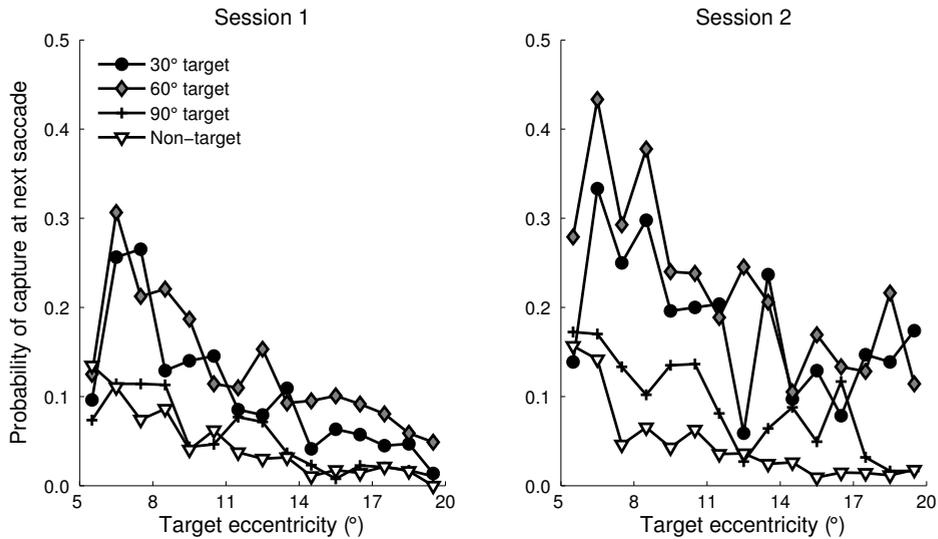


Figure 7.5.: Probability that the next saccade was on target (or comparison non-target) as a function of the distance between fixation and target for session 1 (left panel) and session 2 (right panel).

examined target differences in more detail. Conspicuity areas – the area around the fixation point within which information about the target can be extracted (Engel, 1971; 1977) – were calculated for all three targets according to Motter & Belky (1998a). Figure 7.5 depicts the probability that the target would be hit on a next saccade as a function of target eccentricity at the current fixation. If the 90° target was qualitatively different in terms of target fixation (e.g. because it was more salient), one might expect saccades from further eccentricities than for the 30° and 60° target. As can be seen from Figure 7.5, there was a general decrease of target capture on the next saccade as a function of target eccentricity, for all targets and the mean comparison non-target. There was no qualitatively different pattern for the 90° target. In addition, the 60° target was more often fixated, followed by the 30° target, the 90° target, and finally the comparison non-target. The pattern was similar for session 1 and 2.

### 7.2.2. Characterization of infant saccadic search behavior

In order to characterize infant saccadic search behavior we examined two aspects: fixation durations, and saccade characteristics. As in adult research, we expected the first fixation duration – which is the latency to initiate the first saccade – to be longer than subsequent fixation durations (Hooge & Erkelens, 1996; Zingale & Kowler, 1987). As can be seen from Figure 7.6, the median fixation duration for the first fixation (i.e. the latency to first saccade) was consequently longer than subsequent fixation durations for session 1 and session 2. As fixation durations are not normally distributed (confirmed by separate Kolmogorov-Smirnov tests for first fixation duration and subsequent fixations separately for each session), Wilcoxon rank-sum tests were carried out to assess statistical significance. In session 1, first fixation duration ( $n = 573$ , median = 360.01 ms) was significantly longer than duration of subsequent fixations ( $n = 3525$ , median = 333.24 ms,  $W = 1235801$ ,  $p < 0.05$ ).

In session 2, first fixation duration ( $n = 354$ , median = 359.95) was also significantly longer than duration of subsequent fixations ( $n = 2077$ , median = 313.32,  $W = 496479.5$ ,  $p < 0.05$ ).

To investigate saccade characteristics, the differences in amplitude and direction from one saccade to the subsequent saccade were plotted in a 2-d histogram. Each bin in Figure 7.7 contains those saccades that changed in direction and amplitude compared to the previous saccade as given by the bin edges. For example, if two saccades were made in the same direction, one saccade is added to the bin with the corresponding amplitude change in the left column (no direction change). As visible from Figure 7.7, there are two peaks in the distribution. Saccades generally went in the same direction with equal amplitude as previous saccades or saccades went in the opposite direction with equal amplitude. The latter are saccades that returned to the previously fixated location. This pattern was observed both in session 1 and in session 2.

7. Saccadic search in infancy

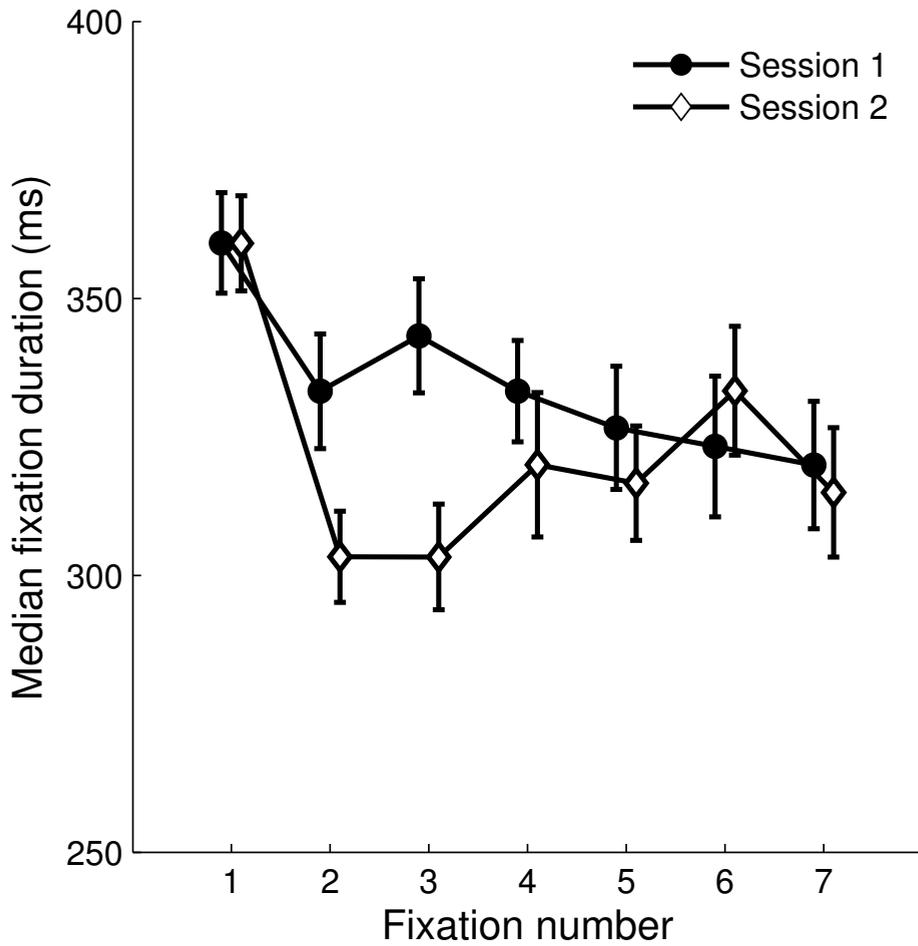


Figure 7.6.: Median fixation duration as a function of fixation number for session 1 and session 2. Error bars depict standard error of the mean. The data between the two sessions are shifted horizontally, as to prevent overlapping error bars.

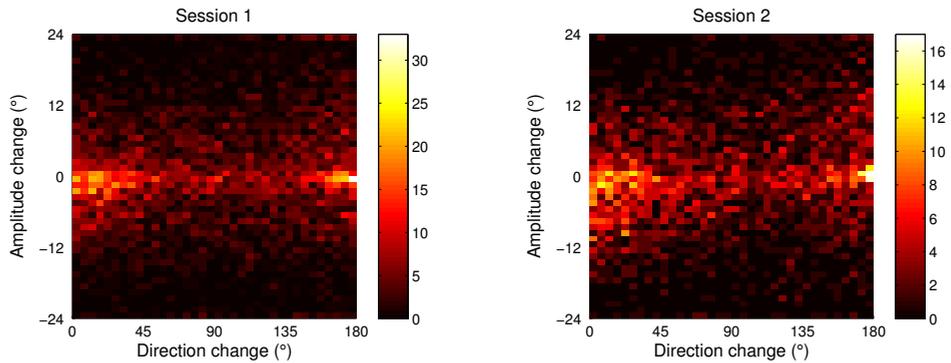


Figure 7.7.: 2-d histograms for the change in amplitude and direction from one saccade to the next for session 1 (left) and session 2 (right). Color indicates frequency of occurrence.

## 7.3. Discussion

Two questions were posed in the present study. First, we questioned whether infants searched for a discrepant item in the absence of instructions. Two components of this question were whether target localization was above what may be expected by chance, and whether target localization was dependent on the target to non-target dissimilarity. Second, we questioned in what manner infants search and we described fixation and saccade characteristics of infant saccadic search. We characterized infant saccadic search behavior in visual search displays, and described how it may relate to adult and primate saccadic search behavior. We presented 10-month-old infants with 24 visual search stimuli and collected eye-movement data in two separate sessions to ascertain the reliability of the findings. All presented findings were consistent across both sessions, unless specifically stated otherwise.

### 7.3.1. Do infants search for a discrepant item in the absence of instructions?

We first report that, depending on the session and target, targets were hit in 30 to 70% of the trials. Infants took on average 1100–1500 ms to

## 7. Saccadic search in infancy

locate the target (approximately 3-4 fixations). Moreover, we noted that saccadic search performance was dependent on the target: saccadic search performance was best for the 60° target, followed by the 30° target, and finally the 90° target. While we initially expected search performance to increase from 30° to 60° to 90° targets, the 90° targets were least likely to be fixated. We return to this point momentarily, but conclude that saccadic search performance was at least target-dependent. Another indication that infants indeed specifically searched for the target, was demonstrated by the 30° and 60° target more readily being fixated than comparison non-targets at equidistant locations from the trial start point (center of the screen). The 90° target was only more readily fixated than comparison non-targets in the second session, not in the first session. We conclude that target fixation in visual search stimuli in infancy is above what can be expected based on our model of chance: in this case the fixation of a non-target bearing no featural differences to the other non-targets. Moreover, the proportion of fixations on a non-target was higher – 22.9% in session 1 and 19.0% in session 2 – than what previous studies expected theoretically (i.e. 12.5% by Amso & Johnson, 2006). To sum, we observed multi-saccadic target localization initiated without instruction, and performance that was target-dependent, both suggesting that infants searched for discrepant items in the absence of instruction.

To substantiate these findings, we examined when infants directed their gaze to the target. If infants searched for the target by looking around until a discrepant item appeared in the periphery, we might expect that the proportion of target-directed saccades increased after the first saccade. If, on the other hand, no active search was taking place, the proportion of target-directed saccades should not change as a function of saccade number. The first is precisely what we observed in the present study. The proportion of target-directed saccades increased after the first saccade, up till around the 3rd or 4th saccade. We suggest that this provides further evidence that infants searched for discrepant items in the absence of instruction.

As explained above, we expected search performance to decrease as a function of target–non-target similarity, yet we observed a different pattern (but consistent across two separate sessions). If the angle between the vertical non-targets and the target is what specifies target–non-target similarity we would expect performance to be best for the 90° target, followed by the 60° and 30° target respectively. The result was that we could not compute linear slopes of visual search performance (i.e. in terms of time to, or number of fixations to, target hit) as a function of target angle. As this was the case, we further investigated how targets were fixated. Specifically, we investigated whether there was a qualitatively different pattern of target fixation as a function of target eccentricity. For example, we might observe that 90° targets were mainly fixated from larger eccentricities compared to the 30° and 60° targets. However, we found that the 90° target was generally fixated less often than the 30° and 60° targets, irrespective of target eccentricity. We therefore suggest that the 90° target may either constitute a less conspicuous target than the 30° and 60° targets, or infants do not consider the 90° line to be a target. This may be due to the 90° target being horizontal, and not slanted as the 30° and 60° target are. If infants categorize slanted objects from horizontal and vertical objects, as has been suggested by earlier research (Quinn & Bhatt, 1998; Treisman & Gormican, 1988), subjective target–non-target similarity for the 90° target is expected to be higher than for slanted objects. Future research into infant visual search performance should benefit from using a smaller range of target angle (i.e. excluding the horizontal target) in order to calculate search slopes as function of target–non-target similarity, or adopt different set sizes to calculate search slopes.

### 7.3.2. In what manner do infants search?

In an attempt to provide a thorough overview of infant saccadic search behavior, we described fixation and saccadic characteristics when infants scan a visual search display. We report that there was a bias for saccades to continue in the direction of the previous saccades with equal amplitude, or to go in the opposite direction with equal amplitude (i.e. saccades to

## 7. Saccadic search in infancy

previously fixated locations). How can these characteristics be interpreted? Previous research on adults and primates has reported on saccade characteristics in visual search as well, and several speculative connections may be made. Hooge et al. (2005), for example, reported that for adults during uniform search (i.e. a task comparable to our present task), as well as during search in pictures and free viewing of pictures, there is also a bimodal distribution of change in saccade direction and amplitude. Saccades generally continued along the current trajectory with the same amplitude or return to the previous fixated location. It appeared as if all elements were systematically investigated, and recently visited locations were sometimes visited again. Infant saccade characteristics are similar to adult saccade characteristics in this regard, although we observed more variation in saccadic direction and amplitude change in infants. In monkey research, on the other hand, Motter & Belky (1998b) reported that saccades in a visual search task have a slight bias to return to the previous fixation location, but otherwise appear directionally random with respect to the previous saccade direction. Future research may focus on determining to what extent human and monkey saccade characteristics are age and stimulus-dependent. Such investigations may inform us whether search strategies for typical real-world search tasks are uniquely human or not, and what role experience plays herein.

In addition to describing infants' saccade characteristics when scanning a visual search display, we examined infants' fixation durations during saccadic search. We expected that the first fixation duration – the latency to initiate a saccade – would be longer than subsequent fixation durations based on adult research (Hooge & Erkelens, 1996; Zingale & Kowler, 1987). We observed this pattern in infants as well. The initial saccade latency is associated with the planning of a sequence of saccades and increases with the number of elements in the task (Zingale & Kowler, 1987). Further research may investigate whether set size affects the saccade latency in infancy as well, indicating when planning of saccade sequences matures. Concluding, there are several notable similarities as well as differences in fixation and saccade characteristics in visual search displays between infants, adults,

and other primates.

### 7.3.3. On the reliability of infant oculomotor characteristics and saccadic search behavior

As most infants in the present study participated twice in the same experiment within two weeks, we had the possibility of assessing test-retest reliability for oculomotor characteristics during saccadic search, and saccadic search performance. While practically all group results described above were similar across both sessions, we wondered how stable individual oculomotor characteristics during saccadic search, and saccadic search performance are. In order to determine test-retest reliability, Pearson's product-moment correlation and intra-class correlation (McGraw & Wong, 1996; Weir, 2005) were calculated for fixation duration, saccade latency, and time to target hit. Figure 7.8 depicts test-retest reliability for oculomotor characteristics during saccadic search: median saccade latency (shown left) and median fixation duration (shown right). Fixation duration showed high test-retest reliability ( $r = 0.71$ ,  $p < 0.05$ ;  $ICC(A,1) = 0.66$ ,  $p < 0.05$ ), while saccade latency was slightly less reliable ( $r = 0.47$ ,  $p < 0.05$ ;  $ICC(A,1) = 0.46$ ,  $p < 0.05$ ).

Fixation duration has previously been shown to be reliable during infant free-viewing (Wass & Smith, 2014), and we extend this finding here to visual search stimuli. Test-retest reliability for saccadic search performance (as determined by time to target hit) was lower for the 60° target than oculomotor characteristics ( $r = 0.47$ ,  $p < 0.05$ ;  $ICC(A,1) = 0.35$ ,  $p < 0.05$ ). Test-retest reliability for the 30° and 90° target was, on the other hand, not reliable at all (both not significantly different from zero). While enough data was available for group comparisons, it is likely there have been too few trials to estimate reliable individual means for saccadic search performance. This is particularly relevant for future studies where individual differences in saccadic search behavior are important, for example in early recognition of Autism Spectrum Disorder (e.g. Gliga et al., 2015). More trials may need to be presented in order to obtain reliable individual estimates of

## 7. Saccadic search in infancy

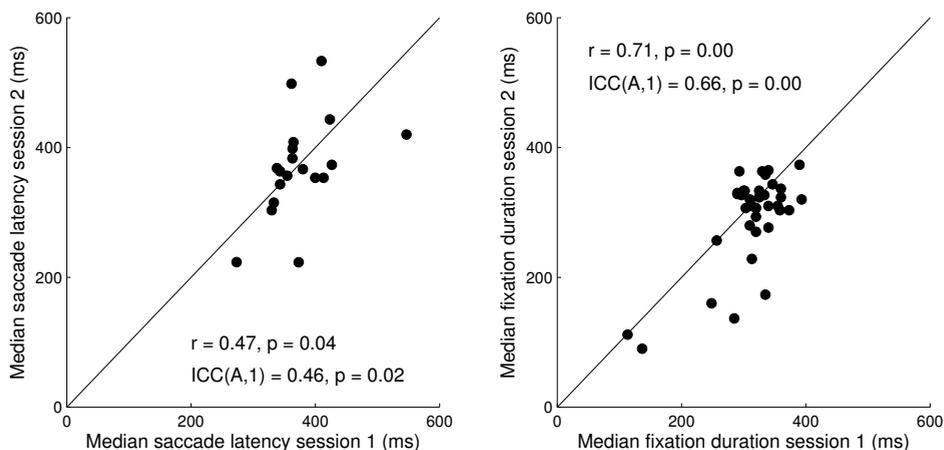


Figure 7.8.: Test-retest reliability for median saccade latency and median fixation duration. Pearson’s product-moment correlation and intra-class correlation are given in text.

saccadic search performance.

## 7.4. Conclusions, limitations, and future research

The main findings of the present study are as follows. Infants searched for discrepant items in the absence of instruction, and saccadic search performance was dependent on target and non-target dissimilarity. Infant saccadic behavior in visual search displays was characterized by saccades following the current trajectory with equal amplitude, and saccades returning to previously fixated locations. These findings were highly consistent across two separate sessions. As saccadic search appears to be above chance in 10-month-old infants, visual search displays are indeed suitable for investigating the development of saccadic search in infancy (Amso & Johnson, 2006; Frank et al., 2014; Schlesinger, Amso, & Johnson, 2007). However, as the empirical chance level was higher than what previous studies theoretically expected, we suggest that future studies always report a proper baseline condition when participants cannot be instructed to perform a visual search task.

#### 7.4. Conclusions, limitations, and future research

While the present study is the first comprehensive description of infant saccadic search behavior, there are several limitations to the study. First, we only measured at one time point (10 months). However, in doing so we have provided a reference point for future research investigating the development of saccadic search performance across the first year after birth. Researchers investigating both the typical and atypical development of visual search may benefit from these results, particularly when investigating the development of saccadic search behavior in Autism Spectrum Disorder. Second, we observed that 90° targets were not more readily detected compared to 30° and 60° targets. In hindsight, using the 90° target prevented us from calculating search slopes as a function of target–non-target similarity. Future research should benefit from using a smaller range of target angles (e.g. 0-45°), or adopting different set sizes to allow search slopes to be calculated.

### **Acknowledgements**

The authors would like to thank all employees at the KinderKennis-Centrum of Utrecht University for help with data collection, and Martijn Schut for proofreading the manuscript. The study was financed through the Consortium on Individual Development (CID). CID is funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant number 024.001.003 awarded to author CK). The funding body had no involvement at any stage of the study.

## References

- Adler, S. A., & Gallego, P. (2014). Search asymmetry and eye movements in infants and adults. *Attention, Perception & Psychophysics*, 76(6):1590–1608.
- Adler, S. A., & Orprecio, J. (2006). The eyes have it: Visual pop-out in infants and adults. *Developmental Science*, 9(2):189–206.
- Amso, D., & Johnson, S. P. (2006). Learning by selection: Visual search and object perception in young infants. *Developmental Psychology*, 42(6):1236–1245.
- Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, 10(4):433–436.
- Caspi, A., Beutter, B. R., & Eckstein, M. P. (2004). The time course of visual information accrual guiding eye movement decisions. *Proceedings of the National Academy of Sciences*, 101(35):13086–13090.
- Colombo, J., Ryther, J. S., Frick, J. E., & Gifford, J. J. (1995). Visual pop-out in infants: Evidence for preattentive search in 3- and 4-month-olds. *Psychological Bulletin & Review*, 2(2):266–268.
- Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review*, 96(3):433–458.
- Engel, F. L. (1971). Visual conspicuity, directed attention and retinal locus. *Vision Research*, 11:563–576.
- Engel, F. L. (1977). Visual conspicuity, visual search and fixation tendencies of the eye. *Vision Research*, 17:95–108.
- Frank, M. C., Amso, D., & Johnson, S. P. (2014). Visual search and attention to faces during early infancy. *Journal of Experimental Child Psychology*, 118:13–26.
- Frank, M. C., Vul, E., & Johnson, S. P. (2009). Development of infants' attention to faces during the first year. *Cognition*, 110(2):160–170.
- Gliga, T., Bedford, R., Charman, T., Johnson, M. H., & The BASIS Team. (2015). Enhanced visual search in infancy predicts emerging autism symptoms. *Current Biology*, 25(13):1727–1730.
- Hessels, R. S., Andersson, R., Hooge, I. T. C., Nyström, M., & Kemner, C. (2015). Consequences of eye color, positioning, and head movement for eye-tracking data quality in infant research. *Infancy*, 20(6):601–633.
- Hessels, R. S., Hooge, I. T. C., Snijders, T. M., & Kemner, C. (2014). Is there a limit to the superiority of individuals with ASD in visual search? *Journal of Autism and Developmental Disorders*, 44(2):443–451.
- Hessels, R. S., Niehorster, D. C., Kemner, C., & Hooge, I. T. C. (2016). Noise-robust fixation detection in eye-movement data: Identification by two-means clustering (I2MC). *Behavior Research Methods*.

#### 7.4. Conclusions, limitations, and future research

- Hooge, I. T. C., & Erkelens, C. J. (1996). Control of fixation duration in a simple search task. *Perception & Psychophysics*, 58(7):969–976.
- Hooge, I. T. C., Over, E. A. B., van Wezel, R. J. A., & Frens, M. A. (2005). Inhibition of return is not a foraging facilitator in saccadic search and free viewing. *Vision Research*, 45(14):1901–1908.
- Hooge, I., & Camps, G. (2013). Scan path entropy and arrow plots: Capturing scanning behavior of multiple observers. *Frontiers in Psychology*, 4:996.
- Johnson, M. H. (1990). Cortical maturation and the development of visual attention in early infancy. *Journal of Cognitive Neuroscience*, 2(2):81–95.
- Kaldy, Z., Giserman, I., Carter, A. S., & Blaser, E. (2016). The mechanisms underlying the ASD advantage in visual search. *Journal of Autism and Developmental Disorders*, 46(5):1513–1527.
- Kaldy, Z., Kraper, C., Carter, A. S., & Blaser, E. (2011). Toddlers with autism spectrum disorder are more successful at visual search than typically developing toddlers. *Developmental Science*, 14(5):980–988.
- Li, Z. (1999). Contextual influences in V1 as a basis for pop out and asymmetry in visual search. *Proceedings of the National Academy of Sciences*, 96:10530–10535.
- McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1):30–46.
- McPeck, R. M., Skavenski, A. A., & Nakayama, K. (2000). Concurrent processing of saccades in visual search. *Vision Research*, 40:2499–2516.
- Motter, B. C., & Belky, E. J. (1998a). The zone of focal attention during active visual search. *Vision Research*, 38(7):1007–1022.
- Motter, B. C., & Belky, E. J. (1998b). The guidance of eye movements during active visual search. *Vision Research*, 38:1805–1815.
- O’Riordan, M. A., Plaisted, K. C., Driver, J., & Baron-Cohen, S. (2001). Superior visual search in autism. *Journal of Experimental Psychology: Human Perception and Performance*, 27(3):719–730.
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):943.
- Quinn, P. C., & Bhatt, R. S. (1998). Visual pop-out in young infants: Convergent evidence and an extension. *Infant Behavior and Development*, 21(2):273–288.
- Schlesinger, M., Amso, D., & Johnson, S. P. (2007). The neural basis for visual selective attention in young infants: A computational account. *Adaptive Behavior*, 15(2):135–148.
- Scinto, L. F. M., Pillalamarri, R., & Karsh, R. (1986). Cognitive strategies for visual search. *Acta Psychologica*, 62:263–292.

## 7. Saccadic search in infancy

- Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12:97–136.
- Treisman, A., & Gormican, S. (1988). Feature analysis in early vision: evidence from search asymmetries. *Psychological Review*, 95(1):15–48.
- Vlaskamp, B. N. S., & Hooge, I. T. C. (2006). Crowding degrades saccadic search performance. *Vision Research*, 46(3):417–425.
- Vlaskamp, B. N. S., Over, E. A. B., & Hooge, I. T. C. (2005). Saccadic search performance: The effect of element spacing. *Experimental Brain Research*, 167(2):246–259.
- Wass, S. V., & Smith, T. J. (2014). Individual differences in infant oculomotor behavior during the viewing of complex naturalistic scenes. *Infancy*, 19(4):352–382.
- Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *Journal of Strength and Conditioning Research*, 19(1):231–240.
- Williams, L. G. (1967). The effects of target specification on objects fixated during visual search. *Acta Psychologica*, 27:355–360.
- Wolfe, J. M. (1994). Guided Search 2.0 A revised model of visual search. *Psychonomic Bulletin & Review*, 1(2):202–238.
- Wolfe, J. M. (1998a). Visual Search. In H. Pashler (Ed.), *Attention*. University College London Press, London, UK.
- Wolfe, J. M. (1998b). What can 1 million trials tell us about visual search? *Psychological Science*, 9(1):33–39.
- Wu, C.-C., & Kowler, E. (2013). Timing of saccadic eye movements during visual search for multiple targets. *Journal of Vision*, 13(11):11.
- Zingale, C. M., & Kowler, E. (1987). Planning sequences of saccades. *Vision Research*, 27(8):1327–1341.

## **Part III.**

# **Gaze behavior to faces in interaction**



## 8. Gaze behavior to faces during dyadic interaction

Published as:

Hessels, R. S., Cornelissen, T. H. W., Hooge, I. T. C., & Kemner, C. (2017). Gaze behavior to faces during dyadic interaction. *Canadian Journal of Experimental Psychology*.

Author contributions:

RH, TC, IH, CK designed the study. TC built the setup. RH collected the data. RH, TC analyzed the data. RH, TC, IH, CK interpreted the data. RH drafted the paper. RH, TC, IH, CK finalized the paper.

## **Abstract**

A long-standing hypothesis is that humans have a bias for fixating the eye region in the faces of others. Most studies have tested this hypothesis with static images or videos of faces, yet recent studies suggest that the use of such “non-responsive” stimuli might overlook an influence of social context. The present study addressed whether the bias for fixating the eye region in faces would persist in a situation that allowed for social interaction. In Experiment 1, we demonstrate a setup in which a duo could engage in social interaction while their eye movements were recorded. Here, we show that there is a bias for fixating the eye region of a partner that is physically present. Moreover, we report that the time 1 partner in a duo spends looking at the eyes is a good predictor of how long the other partner looks at the eyes. In Experiment 2, we investigate whether participants attune to the level of eye contact instigated by a partner by having a confederate pose as one of the partners. The confederate was subsequently instructed to either fixate the eyes of the observer or scan the entire face. Gaze behavior of the confederate did not affect gaze behavior of the observers. We conclude that there is a bias to fixate the eyes when partners can engage in social interaction. In addition, the amount of time spent looking at the eyes is duo-dependent, but not easily manipulated by instructing the gaze behavior of 1 partner.

Imagine the following scenario: you are walking through a crowded square, and each person you walk by looks at you with disgust. You begin to worry – did I spill my morning coffee on my shirt? Is there some toothpaste on my face? By processing information in the faces of others you infer that something is wrong. In this case, the direction of a person’s gaze combined with the emotion portrayed were enough for you to question what you did wrong this morning (see also Gallup, Chong, Kacelnik, Krebs, & Couzin, 2014). But the information contained in faces is not limited to emotion and gaze direction. One may infer someone’s identity from a face, and possibly even someone’s mental state (Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001). Given the amount of information faces can provide, it may be unsurprising that faces tend to attract our attention (Gliga, Elsabbagh, Andravizou, & Johnson, 2009; Langton, Law, Burton, & Schweinberger, 2008) and retain it (Bindemann, Burton, Hooge, Jenkins, & de Haan, 2005). Even newborn infants who are just a few minutes old preferentially follow face-like patterns (Goren, Sarty, & Wu, 1975; Johnson, Dziurawiec, Ellis, & Morton, 1991). Faces are, apparently, something special.

Since the pioneering work of Yarus (1967), researchers have tried to figure out what information humans collect from the faces of others, by investigating eye movements during the exploration of faces. A consistent finding has been that there is a preference for fixating the eye region of faces. This has been suggested to be because of the amount of information that can be extracted solely from the eyes. For example, the eye region is crucial for determining the direction of another’s gaze (Langton, Watt, & Bruce, 2000). The eye region was also most often fixated when participants had to learn a set of faces, supporting the notion of the eyes as primary source of information (Henderson, Williams, & Falk, 2005). Interestingly, Birmingham, Bischof, & Kingstone (2009) reported that this preference for the eyes was irrespective of saliency according to the computational model by Itti & Koch (2000), refuting the idea that facial features are selected purely on the basis of stimulus characteristics.

## 8. Gaze behavior during dyadic interaction

Are the eyes then indeed the primary source of information in faces? V̄o, Smith, Mital, & Henderson (2012) proposed an information-processing account of gaze behavior to faces. They presented videos portraying people in various settings. When the person portrayed looked straight into the camera, there was a preference for fixating the eyes. When the person portrayed spoke, however, the mouth was most often fixated. V̄o et al. (2012) proposed that facial features are fixated based on the amount of information they can provide at a given time. Previous findings in line with this proposal are reported by Birmingham, Bischof, & Kingstone (2008a), who showed that while the eyes of people in social scenes are preferentially fixated, the eyes may be fixated more based on the task required. In their study, people fixated the eyes even more when the direction of people's gaze had to be determined – which, although an intuitive finding, suggests that the context of the scene plays an important role. In a second study, Birmingham, Bischof, & Kingstone (2008b), showed that the social content of the scene may modulate the amount of time spent fixating the eyes. When the number of people increased in the scene, more time was spent looking at the eye regions. Interestingly, when a group of people portrayed active interaction (e.g. playing a game) versus inactivity, the time spent looking at the eye regions increased even more. Social context and task demands thus appear to play an important role in where information is collected in a face.

Despite the apparent importance of social context, most studies have employed pictures of faces or social scenes, thereby ignoring the possible dynamics in gaze behavior when social partners are physically present (Risko, Laidlaw, Freeth, Foulsham, & Kingstone, 2012). An outstanding example of the importance of the physical presence of a social partner was given by Laidlaw, Foulsham, Kuhn, & Kingstone (2011). Their study showed that people frequently looked at a videotaped confederate, but not at a confederate who was actually present in the room. The authors therefore questioned whether there is still a preference for attending the faces of people when a possible interaction partner is physically present. The effect of physical presence on gaze behavior to faces has been further emphasized

by several other recent findings. Freeth, Foulsham, & Kingstone (2013) investigated the effect of social presence on gaze behavior by interviewing participants using either a pre-recorded videotape or a live interviewer. When the interviewer made eye contact, participants looked longer at the face compared with when the interviewer did not make eye contact. Most importantly, this was only the case in the live interaction but not with the pre-recorded video (cf. V̄o et al., 2012), suggesting that physical presence of a social partner matters for gaze behavior to faces and the eyes. Gobel, Kim, & Richardson (2015) furthermore reported that believing that someone might look back modulates the amount of time spent looking at the eyes, based on the social ranking of the inspected person. Gobel et al. (2015) showed participants videos of another person, and the participants were told in separate blocks that they were filmed and that their recording would be shown to that person later on, or that the recording would only be stored in an archive. When participants believed that another person would watch their recording, they spent less time looking at the eyes of people of high social status, and more time looking at the eyes of people of low social status, compared with when the recording was merely stored. Gobel et al. (2015) concluded that gaze has a dual function during social interaction, both retrieving and signaling information, and that “*A complete understanding of face perception needs to address both functions of gaze*” (p. 359). Importantly, such a complete understanding would require investigating gaze behavior to faces in social interaction.

Recently, a new approach to social attention has been proposed, termed Cognitive Ethology (Kingstone, 2009; Kingstone, Smilek, & Eastwood, 2008; Smilek, Birmingham, Cameron, Bischof, & Kingstone, 2006). As the name suggests, this approach advocates studying behavior in social settings as it naturally occurs, before moving to the laboratory instead of vice versa. This approach may be particularly useful in studying the dual function of gaze as suggested by Gobel et al. (2015). One such study has recently been conducted by Jarick & Kingstone (2015). Participants were either instructed to cooperate in duos on a tangram puzzle, or had to compete against each other. Hereafter, each duo was asked to maintain eye

## 8. Gaze behavior during dyadic interaction

contact for 10 minutes. Duos who just competed against each other on the puzzle maintained eye contact for a significantly longer time, and smiled, talked and laughed for a significantly shorter time compared with duos that just cooperated on the puzzle. From this, Jarick & Kingstone (2015) conclude, in accordance with Gobel et al. (2015), that “... *when looking at the eyes of a real person one both acquires and signals information to the other person*” (p. 7). This is corroborated by Ho, Foulsham, & Kingstone (2015), who reported that the eyes can signal the beginning and end of one’s turn in conversation. While Jarick & Kingstone (2015) demonstrate a striking difference in behavior after just a short competitive or cooperative game, we were left with a number of questions. While participants were asked to make eye contact while monitored by the experimenter (and recorded on video for later coding), there was no objective measurement that participants indeed looked at the eyes of the other. Where do people look when they (are instructed to) look towards each other? Is there still a bias for fixating the eye region when there is the possibility of social interaction (cf. Laidlaw et al., 2011)? We address these questions in the present study.

Following the reasoning of the cognitive ethology approach addressed earlier, we aimed to allow social interaction while still maintaining a degree of control in a lab environment. Social interaction is a broad term, we define it here as exchange of behavior between two partners, be it by conversing, laughing, or interaction in gaze behavior. More specifically, we wanted to investigate gaze patterns to faces during social interaction. We therefore built a two-way video setup, capable of recording eye movements from two participants while they looked at each other and had the possibility of engaging in interaction. Participants were placed behind a two-way video setup in the same room. Although this did impose a physical barrier between participants, it allowed us to investigate eye-movements to faces with high resolution without placing eye trackers on somebody’s head (i.e. by using mobile eye trackers), and without the geometric problems of eye-tracking a 3-d world instead of a 2-d plane. To our knowledge, this is the first eye-tracking study using a setup built for high-resolution eye tracking of gaze behavior to facial features in social interaction. Our first research

question pertains to the bias for attending the eyes: is there still a bias for attending the eyes when there's possibility of interaction? If a bias to fixate the eyes is only present in the absence of social partners, we should not find a preference for fixating the eyes in the current experiment. Our second question relates to the gaze behavior of two people looking at each other: is the gaze pattern of one partner predictive of the gaze pattern of the other partner? If maintaining eye contact reveals something about the nature of the relationship between the two interaction partners, for example competitive versus cooperative (Jarick & Kingstone, 2015), we expect gaze patterns to be linked to each other. More specifically, is there interaction between the two partners in a duo with regard to looking at the eyes, such that the time spent looking at each other's eyes is correlated?

### 8.1. Method experiment 1

#### 8.1.1. Participants

Twenty-two individuals volunteered in the first experiment of the study. All participants were students or employees at the Faculty of Social and Behavioral Sciences at Utrecht University, the Netherlands. Of the 22 participants, 4 were excluded because of technical difficulties. More specifically because of a wrong settings file ( $n = 1$ ), insufficient lighting ( $n = 2$ ), and calibration problems ( $n = 1$ ). Of the 18 remaining participants, 8 were male. Mean age was 30 years ( $sd = 8$  years). The 18 participants were arranged into pairs for the experiment: 3 male-male pairs, 4 female-female pairs, and 2 male-female pairs. All participants gave written informed consent prior to the start of the experiment.

#### 8.1.2. Apparatus & stimuli

Figure 8.1 depicts a schematic overview of the two-way video setup used in the present study (see Appendix for pictures of the setup). Participants were seated in the same room in front of a wooden box at either end of an office table of approximately 2 meters length. Inside this wooden box was a half-silvered mirror, which reflected the screen lying in the bottom of

## 8. Gaze behavior during dyadic interaction

the box. If participants did not look down, it would be hard to notice that they were in fact not looking directly at a screen. Behind the half-silvered mirror a webcam was placed. The webcam could record video of the participant through the mirror. This video image was subsequently presented on the screen of the other participant, which reflected from the mirror. Two Logitech webcams recorded video frames of 800 by 600 pixels at 30 Hz. A stimulus computer running Ubuntu 12.04 LTS, Spyder 2.1.9, Python 2.7.3, NumPy 1.6.1, and OpenCV 2 was used for the presentation of the videos on two 22" Dell P2213 screens. Video frames were scaled to a resolution of 1024 by 768 pixels, and were presented in the center of the 1680 by 1050 pixels available on the screen. The video frames were surrounded by a black border. Videos were concurrently presented to the other participant and recorded to disk. Two SMI RED eye-tracking systems running at 120 Hz were used for the collection of eye-movement data during the presentation of the videos. Two Windows 7 computers running iViewX 2.8 build 26 controlled the SMI RED systems. A parallel port connection from the Ubuntu stimulus computer to the two eye-tracker computers signaled the start and stop of the eye-tracking recording, and the start and stop of the two-way video stream.

Latency from webcam to screen was measured using a Casio Exilim EX-ZR200 1000 Hz camera, and was on average 119 ms ( $sd = 14.8$  ms). Distance from the participants' eyes to the screen was 81 cm (see the dashed lines in Figure 8.1). However, as the video image of the participant is not true to size, the visual viewing distance from participant to participant was not identical to the physical viewing distance (i.e. twice 81 cm). The visual viewing distance was calculated by placing an object of known size at one end of the setup and calculating its visual angle in degrees from the other side of the setup. The visual viewing distance was calculated to be 136 cm; in other words, participants viewed each other through the video setup as if they were seated at that distance without the video setup. The 136 cm visual viewing distance was roughly a meter less than the physical distance between the participants in the room. As participants were in the same room, they could hear each other when either participant talked. Previous

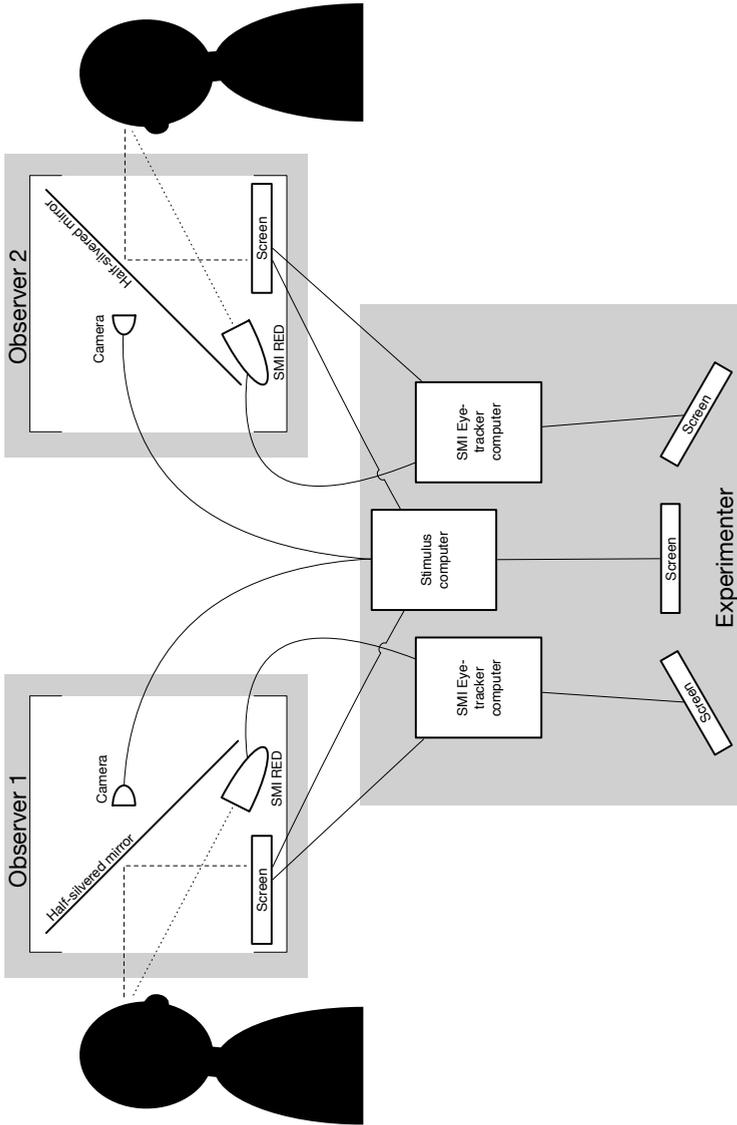


Figure 8.1.: Schematic overview of the experimental setup. The observer part is depicted from the side, the experimenter part from the top.

## 8. Gaze behavior during dyadic interaction

research indicates that when the audio signal precedes the video signal in speech integration the minimum lag needed to notice a lag is around 130 ms (Dixon & Spitz, 1980). The latency between webcam and screen (119 ms, as noted above), combined with the time it takes the audio to travel from one participant to another (roughly 7 ms based on the speed of sound traveling 2.36 m at room temperature) means the audio-video latency was roughly 110 ms. This is 20 ms below the just noticeable difference reported in earlier research. Moreover, this did not result in a subjectively noticeable lag during pilot studies.

### 8.1.3. Procedure

After informed consent was given, participants were positioned in front of one of the two parts of the two-way video setup. Participants were positioned such that the distance from the participants' eyes to the eye tracker was 70 cm (see the dotted lines in Figure 8.1), and the eyes were at the same height as the webcams. The latter was done to ensure that when participants looked straight ahead they looked into the camera. Moreover, the screen was positioned such that this would correspond to eye height of the other participant. I.e. when both participants look straight ahead they look into each other's eyes, and straight into the camera. Following positioning, a 5-point calibration was run from the eye-tracker computers. After calibration, a 4-point validation was run. We aimed to minimize the average offset between point of regard and the validation points. All participants had an average offset value that was at least below  $1.25^\circ$ . Only one participant was excluded because of calibration problems.

Once both participants were positioned and calibrated, instructions on the experiment were given. Participants were instructed that they were to look at each other for 5 minutes (similar to Jarick & Kingstone, 2015, but without the precise instruction on eye contact). If they failed to do so for whatever reason, they were to continue as soon as possible and finish the rest of the 5 minutes. The experimenter notified the participants once the 5 minutes were up. No explicit instructions about where to look were

given (i.e. both maintaining eye contact or scanning the entire face was perfectly allowed), and no restrictions were given with respect to talking or laughing. As few instructions as needed were given, to refrain from focusing participants on their own (gaze) behavior. This was done to allow participants to behave in whatever way they felt natural in this 5-minute period as long as they looked towards the screen on which the other participant was displayed and not the rest of the room.

#### **8.1.4. Data reduction**

Eye-movement data were exported to text files, imported into MATLAB 2014a, and subsequently trimmed to the start and end of the videos. Hereafter, eye-movement data of each participant were linked to the video of the other participant by means of the following steps:

1. Using custom MATLAB software, Areas of Interest (AOI) centers were determined for the left eye, right eye, nose, and mouth. The experimenter first selected the center of the pupils, the tip of the nose, and the center of the mouth. Hereafter, face-tracking software tracked the location and orientation of the participant's face through the video and updated the location of the eyes, nose, and mouth for each video frame. The main advantage of the present method for the construction of AOIs is that it is semi-automatic, and requires significantly less time as compared with manual frame-by-frame AOI construction in videos. After selecting the AOI centers, the software ran automatic. If the AOI center as detected by the software would gradually move of the physical AOI center, the experimenter could intervene. The total time the experimenter is involved with the software amounts to about a minute per video. Manual coding is usually done frame-by-frame, taking at least a multiple of the video duration.
2. As the eye-movement data were recorded at 120 Hz, and the video only at 30 Hz, eye-movement data were down sampled using a moving-window average to the frame rate of the video. Moreover, down sampling increases the signal-to-noise ratio of the eye-movement data by

## 8. Gaze behavior during dyadic interaction

a factor of the square root of the number of samples averaged (the square root of 4 in this case). Down-sampled data were subsequently combined with the AOI information acquired in step 1.

3. For each video frame, the corresponding down-sampled gaze coordinate of the participant was assigned to one of four AOIs – left eye, right eye, nose, and mouth – using the Limited-Radius Voronoi Tessellation (LRVT) method (Hessels, Kemner, van den Boomen, & Hooge, 2016). The LRVT method works by computing the Euclidean distance of the gaze coordinate to all AOI centers. Hereafter, the AOI that has the cell-center closest in space to the gaze coordinate is assigned to this sample, but only if it does not exceed the Limited Radius. As faces are sparse stimuli and gaze is clustered on the facial features, the size of the AOI does not affect the outcome measures (e.g. total dwell time) beyond a certain size (Hessels et al., 2016). The LRVT method with large radii, and the Voronoi method from which it derives have been shown to be most robust to noise in face stimuli. The radius was set to  $4.0^\circ$ . In comparison, the average eye-to-eye distance across all videos was roughly  $3^\circ$ , and the mean distance from each AOI to its closest neighbor (or AOI span; see Hessels et al., 2016) was around  $2^\circ$ .
4. Subsequent video frames for which gaze was at the same AOI were merged into dwells. As the dwell analysis was done sample based as opposed to fixation based, it may be that the AOI signal contains very short dwells because of noise instead of two saccades to and from another facial feature (see Holmqvist et al., 2011, p. 223–224 for a discussion). Dwells of one video frame (33.3 ms) surrounded by dwells on the same AOI – for example left eye, nose, left eye – were therefore excluded and the surrounding dwells merged.
5. Total dwell times (Holmqvist et al., 2011, p. 389) to the eyes, nose, and mouth were calculated by summing the durations of all dwells to the corresponding AOI. The total dwell times to the left and right eye were summed for the eyes AOI. Dwell times (Holmqvist et al., 2011, p.

386–389) were calculated by averaging the individual dwells on eyes, nose, and mouth. A non AOI was used for all frames in which gaze was available, but not on any of the three other AOIs; this included all gaze data on or off the screen, but not lost data. Finally, the average duration of eye contact, as defined by simultaneous dwells on the eyes AOI by both participants in a duo, was calculated.

## 8.2. Results experiment 1

### 8.2.1. Total dwell time

To determine whether there was a bias for fixating the eyes, a repeated-measures ANOVA on total dwell times with the factor AOI (eyes, nose, mouth and non AOI) was used. As the assumption of sphericity was violated, F-values were corrected using Greenhouse-Geisser corrections. The effect of AOI was significant ( $F(1.43,24.25) = 21.51, p < .001, \eta^2 = 0.56$ ). Post-hoc paired-sample t-tests with a Bonferroni-corrected alpha of .008 revealed that total dwell time on the eyes ( $m = 151.05, sd = 81.62$  seconds), was significantly longer than the total dwell time on the nose ( $m = 53.90, sd = 39.60$  seconds), mouth ( $m = 30.51, sd = 43.16$  seconds), and non AOI ( $m = 11.48, sd = 14.87$  seconds), all  $p < .001$ . The total dwell time on the nose was significantly longer than the total dwell time on the non AOI. The differences between the total dwell time on the nose and the mouth ( $p = 0.089$ ), and the mouth and the non AOI ( $p = 0.028$ ) were non-significant at the Bonferroni-corrected alpha of .008. Total dwell times are given in Figure 8.2.

Intra-class correlations (McGraw & Wong, 1996; Weir, 2005) were calculated to examine whether total dwell times on the eyes, nose, mouth and non AOI were correlated between the participants in each duo. The intra-class correlations for absolute agreement were calculated, so that a perfect correlation means the same amount of time spent looking at the AOI, i.e. perfect reciprocity in the total dwell time to an AOI. A significant correlation for the eyes AOI ( $ICC[A,1] = 0.75, p < .005$ ) indicated a strong

8. Gaze behavior during dyadic interaction

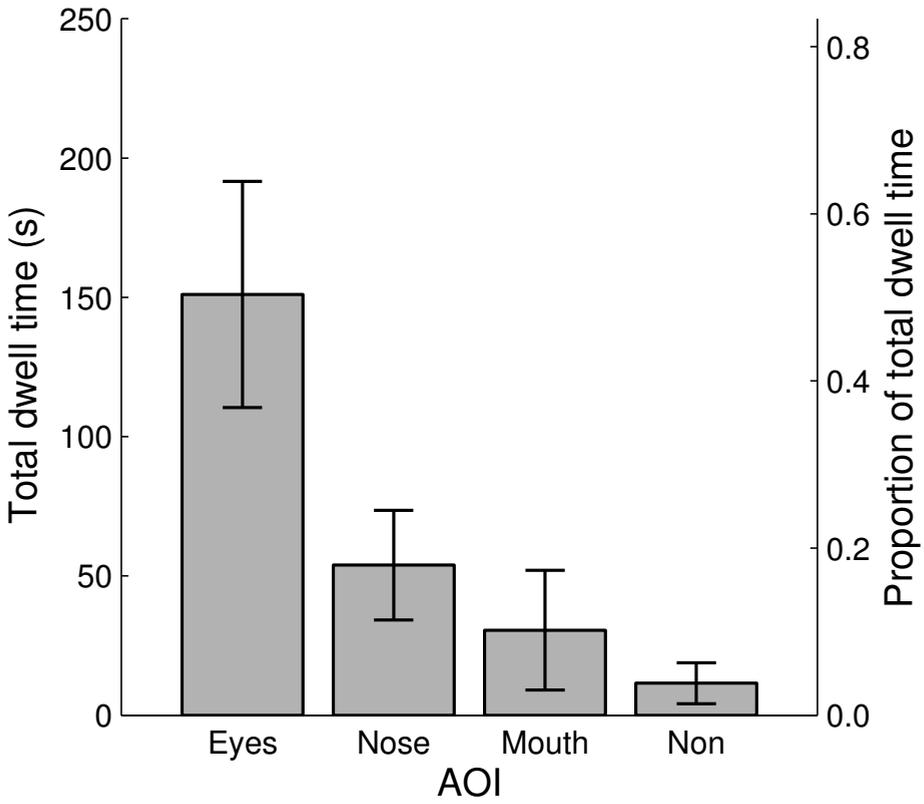


Figure 8.2.: (Proportion of) total dwell time to all four AOIs in experiment 1. Error bars depict 95% confidence interval.

relation between the time one participant in a duo spent looking at the eyes and the time spent looking at the eyes by the other participant in the duo. Correlations for the nose, mouth, and non AOI were non-significant (all  $p > .1$ ). Scatter plots for total dwell time to all four AOIs are given in Figure 8.3.

### 8.2.2. Dwell time and eye contact

In order to get a better understanding of gaze behavior over time, we calculated dwell times. The dwell time refers to the average time gaze stays in one AOI. A repeated-measures ANOVA on dwell times with the factor AOI (eyes, nose, mouth and non AOI) was run. As the assumption of sphericity was violated, F-values were corrected using Greenhouse-Geisser corrections. The effect of AOI was significant ( $F(2.07,28.96) = 3.84, p < .05, \eta^2 = 0.22$ ). Post-hoc paired-sample t-tests with a Bonferroni-corrected alpha of .008 revealed that dwell time on the eyes ( $m = 0.80, sd = 0.40$  seconds), was significantly longer than the dwell time on the nose ( $m = 0.48, sd = 0.25$  seconds) and mouth ( $m = 0.45, sd = 0.17$  seconds), both  $p < .008$ . No other differences between dwell times to the eyes, nose, mouth and non AOI were significant (all  $p > .01$ ). The average duration of eye contact, as defined by both partners in a duo looking at the eye AOI simultaneously, was 0.43 seconds ( $sd = 0.19$ ). Dwell times and the duration of eye contact are given in Figure 8.4.

### 8.2.3. Descriptive analysis of gaze behavior over time

Scarf plots of the dwells to all four AOIs (eyes, nose, mouth, and non AOI) for five duos are depicted in Figure 8.5 (see Appendix for remaining duos). These five were chosen for the large differences in the total dwell time to the eyes. The panels A-E are ordered by the total dwell time to the eyes. For example, the first pair (panel A) not only spent time on the eyes AOI, but also on the nose and mouth AOIs, whereas the fourth and fifth pair (panels D-E) spent most of the time on the eyes AOI with only a few dwells on the nose AOI. There are several key findings that can be drawn from these scarf plots. First, as already noted in the total dwell

8. Gaze behavior during dyadic interaction

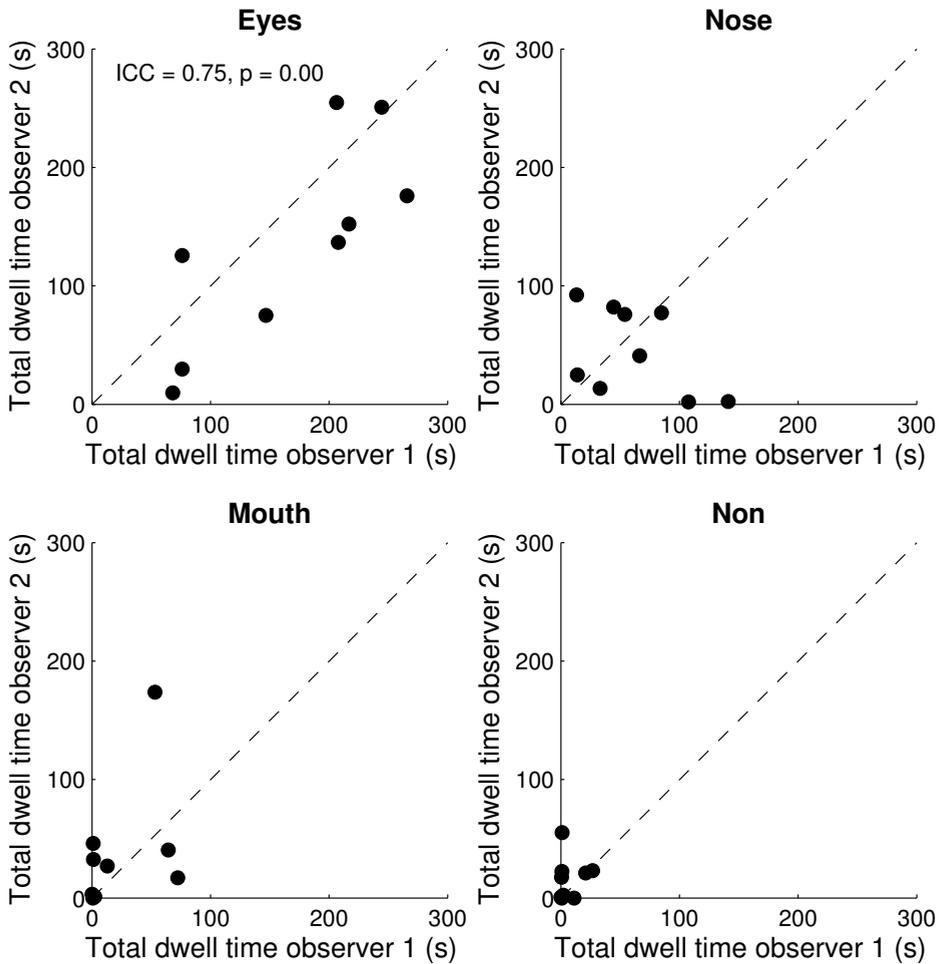


Figure 8.3.: Correlations between total dwell time to the eyes (top left), nose (top right), mouth (bottom left), and non AOI (bottom right) for observer 1 and observer 2 in experiment 1. Significant intra-class correlations are given in text. All points falling on the unity line would indicate a perfect correlation between the gaze behaviors of the two observers.

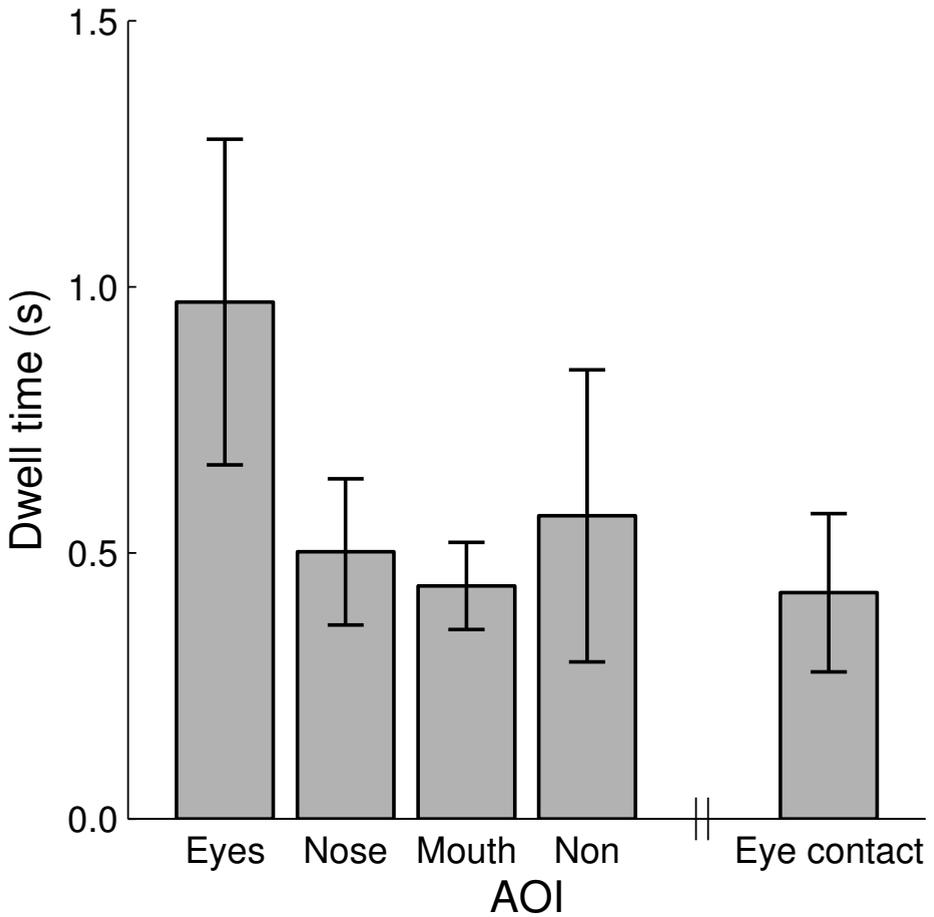


Figure 8.4.: Dwell times to all four AOIs, and the average duration of eye contact (both partners looking at the eyes AOI simultaneously) in experiment 1. Error bars depict 95% confidence interval.

## 8. Gaze behavior during dyadic interaction

time analysis, participants varied greatly in the amount of time they spent on the eyes AOI. Second, as noted in the dwell time analysis, dwells were generally quite short; on the order of hundreds of milliseconds to a few seconds. Third, different AOIs may be looked at in quick succession. For example, observer 1 in panel C looked at the eyes, nose, and mouth AOI a large number of times within the first 50 seconds.

In addition to the general patterns that emerge, there are a number of specific scenarios that occurred for the duos. For example, observer 1 in panel A started by looking at the eyes of observer 2 for a couple of seconds. Hereafter, dwells were mainly on the mouth and non AOI (i.e. outside the facial features), with only a few brief dwells on the eyes. It appears as if observer 1 was avoiding the eyes of observer 2 after an initial dwell on the eyes in the first seconds of the experiment. For the duo in panel C, another interesting pattern emerged. Observer 1 in this duo looked at the eyes, nose, and mouth repeatedly in the first half of the experiment, yet looked mainly at the eyes in the second half of the experiment. For observer 2, this pattern seems reversed. In the first half of the experiment, observer 2 mainly looked at the eyes, whereas gaze was divided between the eyes, nose and non AOI towards the second half of the experiment. It would appear as if the gaze behavior of observer 1 is inversely related to the gaze behavior of observer 2 for this duo. If we compare this to the duo in panel E, we observe a different pattern. Here, observer 1 and 2 started the experiment with dwells mainly on the eyes, combined with a few dwells on the nose, mouth and non AOI. Towards the second half of the experiment, both observers mainly dwelled on the eyes, with only a few brief dwells on the nose. Here it would seem that the gaze behaviors of the two observers were more similar across time to each other compared with the observers in panel C. To sum, the exact relation between the gaze behavior of the two partners appears to be duo-dependent.

### 8.3. Discussion experiment 1

Duos were asked to look towards each other for 5 minutes while their eye movements were being recorded. Two questions were posed for the first experiment. 1) Is there still a bias for attending the eyes when there's possibility of interaction? 2) Is the gaze pattern of one partner predictive of the gaze pattern of the other partner? For the first research question, we report longer total dwell times to the eyes compared with the nose, mouth and non AOI and conclude that there is a bias for fixating the eye region during social interaction, as has often been observed in non-interaction settings. Moreover, the average dwell duration to the eyes was longer than the dwell durations to the nose and mouth. Periods of eye contact took on average less than 0.5 seconds. We will discuss the implications hereof in the General discussion. For the second research question, we report that looking at the eyes appears coupled: total dwell time to the eyes was highly correlated between participants. When one participant generally looked for a long total duration at the eyes for example, the other participant in the duo did so too. While the total time spent looking at the eyes by one interaction partner seems predictive of the total time spent looking at the eyes by the other partner, the relationship between the two across time seems to be duo-dependent.

If the correlation between total dwell times to the eyes is a result of two partners attuning to the level of eye contact instigated by either partner, changing the gaze behavior of one partner should result in a change in the gaze behavior of the other. More specifically, if one partner looks longer into the eyes of the other, the second partner should follow. This is precisely what we investigated in experiment 2 by recruiting a confederate to participate as one of the partners in each duo, and guiding that confederate's gaze behavior.

## 8. Gaze behavior during dyadic interaction

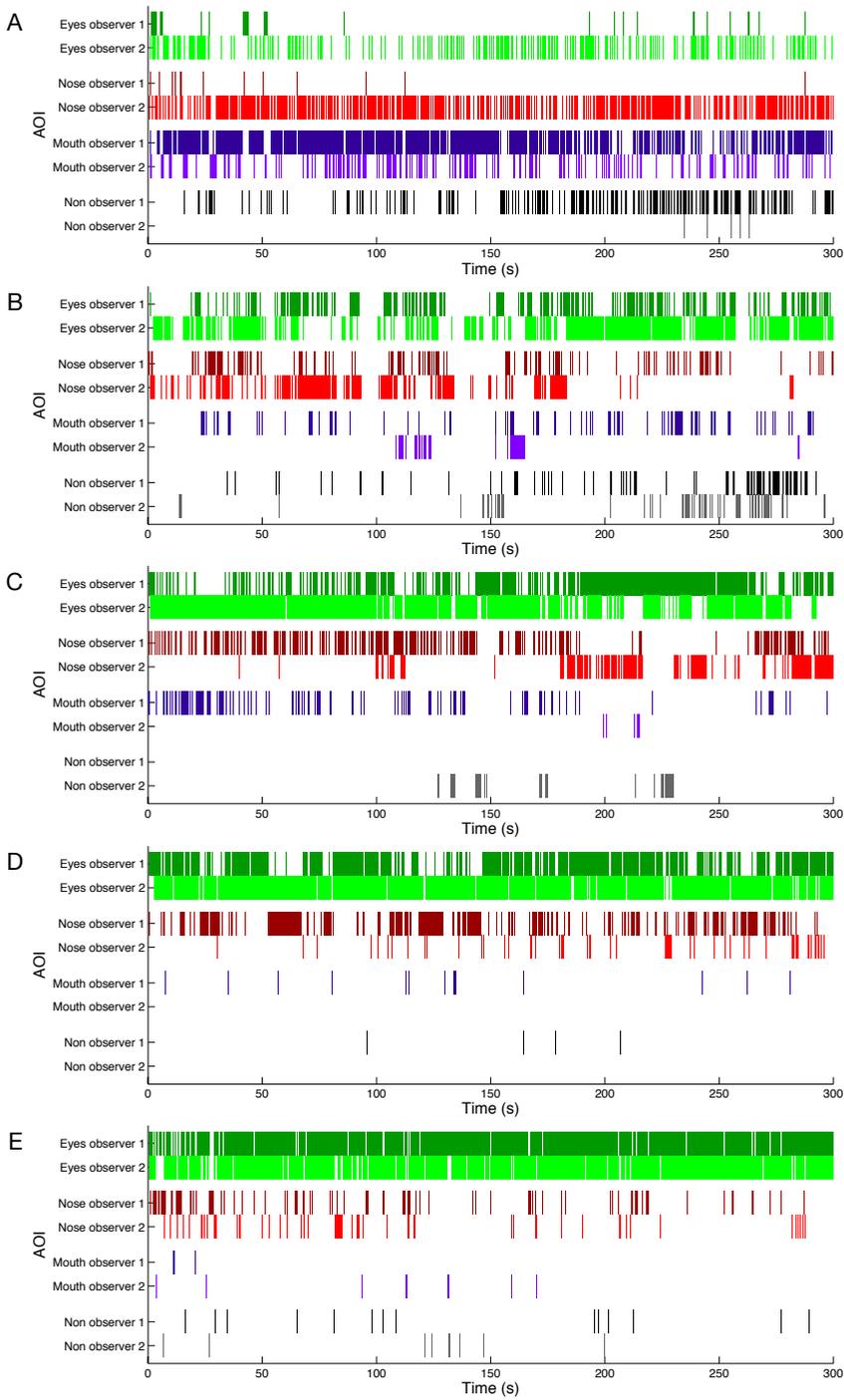


Figure 8.5.: Scarf plots of dwells to the four AOIs for 5 duos (panels A-E) in order of the total dwell time to the eyes in experiment 1. Each filled rectangle represents one dwell with its length equal to the duration of that dwell. The dark colored bars in each scarf plot belong to observer 1, whereas the lighter colored bars belong to observer 2. Bars are grouped by each AOI (eyes, nose, mouth, and non AOI). This allows easy comparison of dwells to the AOIs as a function of time between participants.

## 8.4. Method experiment 2

### 8.4.1. Participants

Thirty-seven individuals volunteered in the second experiment of the study. All were students, visitors or employees at the Faculty of Social and Behavioral Sciences of Utrecht University, the Netherlands. When students were eligible, course credit was provided as compensation of participation. Of the 37 participants, 7 were excluded because of technical difficulties ( $n = 2$ , i.e. problems with the instructions of the confederate), poor data quality ( $n = 4$ ), and calibration problems ( $n = 1$ ). Of the 30 remaining participants, 15 were male, 15 were female. Mean age was 26 years ( $sd = 4$  years). One confederate (24 year-old male) was recruited to assist as one of the interaction partners in each measurement. We henceforth refer to observers to avoid confusion with the term participants, which included the confederate as well. All participants gave written informed consent prior to the start of the experiment.

### 8.4.2. Apparatus, stimuli & procedure

The apparatus, stimuli & procedure were identical to experiment 1 with the exception of the instructions the confederate received. The confederate was fitted with an earpiece attached to an iPod through which pre-recorded instructions were given. A beanie (i.e. winter hat) was used to cover the earpiece without occluding the facial features. As the measurements were done in the winter, we reasoned a beanie would not stand out as abnormal. Moreover, debriefing revealed that no observer suspected that the confederate was given instructions.

We expected that if one partner looks longer into the eyes of the other, the second partner should follow. Hence we instructed the confederate to either fixate the eyes of the observer for 150 seconds straight, or scan the face of the observer for 150 seconds by looking at consecutively instructed locations on the observer's face. The locations that were fixated when the confederate scanned the observer's face were the left eye, right eye, nose,

mouth, left ear, right ear, left cheek, right cheek, chin, and forehead. One sequence of random successive locations was pre-recorded. The time between locations was varied by sampling from a normal distribution with a mean of 4 seconds, and standard deviation of 0.3 seconds. Observers were randomly assigned to one of two groups when entering the lab. 15 observers were included in each group, and for each group the confederate received the instructions in different order. For the first group the confederate would look at the left eye of the observer for 150 seconds, followed by 150 seconds of scanning the face (i.e. as determined by the pre-recorded instructions). This order will be referred to as *fixate eyes instruction first*. For the second group the confederate would scan the entire face first for 150 seconds, followed by 150 seconds of looking at the left eye of the observer. This order will be referred to as *scan face instruction first*. The reason for instructing the participant to fixate the left eye, as opposed to, for example, the nose bridge exactly between the eyes, was because the confederate could then use the pupil of that eye as a physical feature to retain focus on. The nose bridge between the eyes is a less well-defined location, and may cause the confederate to accidentally start looking at the nose. The pre-recorded instructions were switched for use in both groups, i.e. the first part of the *scan face instruction first* was identical to the second part of the *fixate eyes instruction first*. The confederate was instructed not to initiate conversation, but to respond to bids for conversation as he normally would.

### 8.4.3. Data reduction

Identical to experiment 1 with the exception that total dwell times were also calculated separately for the first 150 seconds and last 150 seconds of the experiment.

## 8.5. Results experiment 2

### 8.5.1. Total dwell time

Before we examined whether observers' gaze behavior was dependent on the gaze behavior of the confederate, we first examined whether total dwell

## 8. Gaze behavior during dyadic interaction

times to the facial features mimicked what we observed in experiment 1 using a repeated-measures ANOVA on total dwell times with the factor AOI (eyes, nose, mouth and non AOI). As the assumption of sphericity was violated, F-values were corrected using Greenhouse-Geisser corrections. The effect of AOI was significant ( $F(1.37,39.77) = 43.03, p < .001, \eta^2 = 0.60$ ). Post-hoc paired-sample t-tests with a Bonferroni-corrected alpha of .008 revealed that total dwell time on the eyes ( $m = 152.42, sd = 71.53$  seconds), was significantly longer than the total dwell time on the nose ( $m = 64.79, sd = 41.99$  seconds), mouth ( $m = 32.58, sd = 31.59$  seconds), and non AOI ( $m = 11.54, sd = 16.43$  seconds), all  $p < .001$ . Moreover, the total dwell time on the nose was significantly longer than the total dwell time on the mouth and non AOI (both  $p < .001$ ). Finally, the total dwell time on the mouth was significantly longer than the total dwell time on the non AOI ( $p < 0.005$ ). Total dwell times are given in Figure 8.6. In conclusion, in experiment 2 there was also a bias for fixating the eyes of the confederate, and total dwell times across the facial features were highly comparable to those obtained in experiment 1.

### 8.5.2. The effect of confederate gaze behavior – total dwell time to the eyes

To confirm that the confederate followed the instructions to fixate the eyes or scan the entire face, a repeated-measures ANOVA on the confederate's total dwell time to the eyes was run with confederate instructions (fixate eyes, scan face) as a within-subject factor and order (*fixate eyes instruction first, scan face instruction first*) as a between-subject factor. A significant main effect of confederate instructions ( $F(1,28) = 272.42, p < .001, \eta^2 = 0.91$ ) revealed that the confederate looked longer at the eyes when he was instructed to do so compared with when he was instructed to scan the face. The main effect of order was non-significant ( $F < 1$ ), as was the interaction between confederate instructions and order ( $F < 1$ ). As visible from the left panel of Figure 8.7, the confederate looked at the eyes of the observer for roughly 130 seconds when he was instructed to fixate the eyes, and only roughly 80 seconds when he was instructed to scan the face.

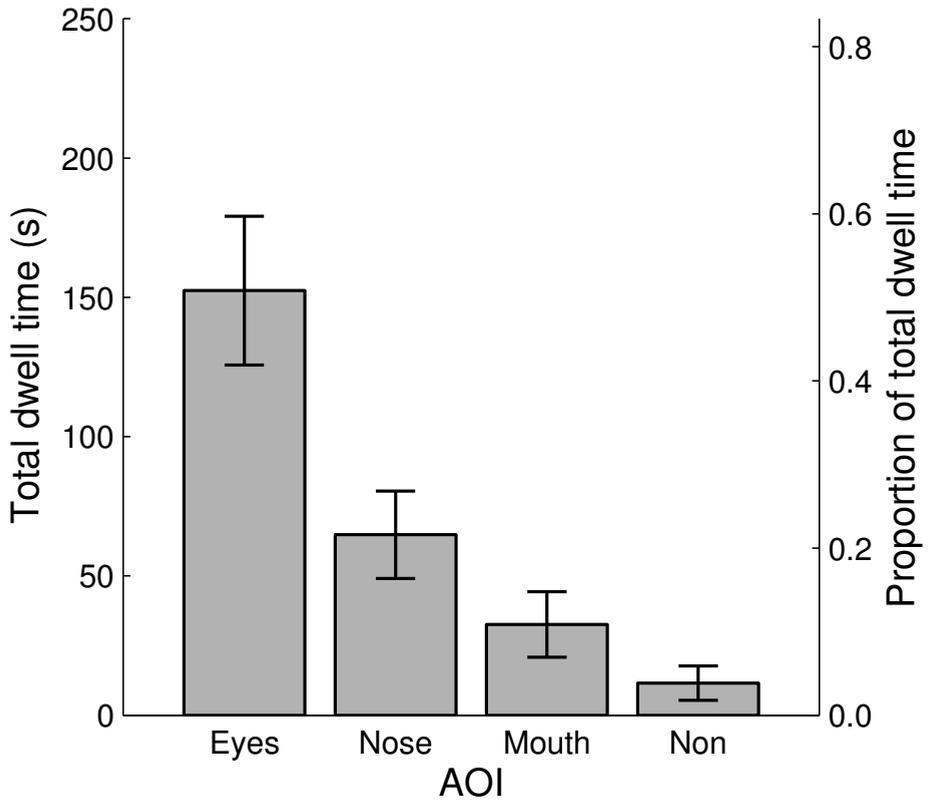


Figure 8.6.: (Proportion of) total dwell time to all four AOIs in experiment 2 by the observers (i.e. confederate excluded). Error bars depict 95% confidence interval.

## 8. Gaze behavior during dyadic interaction

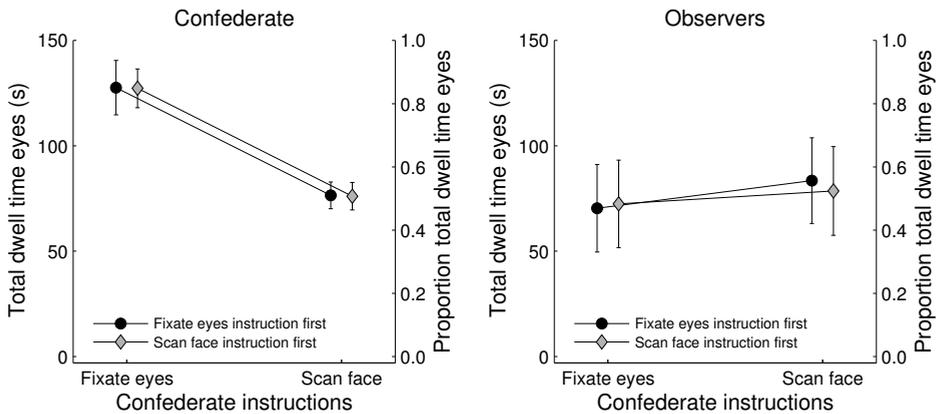


Figure 8.7.: (Proportion of) total dwell time to the eyes for the confederate (left panel) and observers (right panel) as a function of the instructions given to the confederate in experiment 2. Separate lines depict the order in which the confederate fixated the eyes and scanned the face. A horizontal shift for one of the lines has been introduced on the x-axis to facilitate interpretation of the graph. Error bars depict 95% confidence interval.

This indicates that the confederate followed the instructions (see Appendix for additional analyses supporting this conclusion).

To investigate whether the gaze behavior of the confederate elicited a change in the time the observers fixated the eyes of the confederate, a repeated-measures ANOVA on observers' total dwell time to the eyes was run with confederate instructions (fixate eyes, scan face) as a within-subject factor and order (*fixate eyes instruction first*, *scan face instruction first*) as a between-subject factor. If the observers' gaze behavior was indeed coupled to the gaze behavior of the confederate, we would expect longer total dwell times to the eyes for the observers when the confederate fixates the eyes of the observer. More specifically, we would expect the pattern we saw for the confederate in Figure 8.7 (left panel) to emerge for the observers as well. As visible from Figure 8.7 (right panel), this was not the case. Total dwell times to the eyes were similar regardless of confederate

instructions, or the specific group participants were assigned to. Indeed, a non-significant main effect of confederate instructions ( $F(1,28) = 3.18, p > .05$ ) revealed that observer's total dwell time to the confederate's eyes did not differ based on the confederate's gaze behavior. In addition, the order in which the confederate fixated the eyes and scanned the face of the observer did not affect observers' total dwell time to the confederate's eyes ( $F < 1$ ). Finally, there was no interaction of order and confederate instructions ( $F < 1$ ). To sum, the observers' total dwell time to the confederate's eyes did not depend on how long the confederate looked at the observers' eyes nor when he did so.

### **8.5.3. The effect of confederate gaze behavior – entropy of total dwell times**

To ascertain that no changes in the observers' gaze behavior were elicited by the confederate's gaze behavior, two alternatives were explored. First, gaze behavior to the total face was considered. While observers' total dwell time to the confederate's eyes was not dependent on whether the confederate fixated the eyes or scanned the face of the observers, it may be that observers distributed their time more equally over the rest of the facial features when the confederate scanned the face. For example, it may be that the observers divided gaze between the eyes and nose when the confederate fixated the eyes, but that they divided gaze between eyes, nose, mouth, and other areas of the screen when the confederate scanned the face. In other words, did the observer scan the face more when the confederate did so too, as compared with when the confederate fixated just the eyes? To determine this, Shannon entropy (Shannon, 1948), a measure of uncertainty, was calculated over the proportion of total dwell time to the left eye, right eye, nose, mouth and non AOI (see e.g. Hooge & Camps, 2013). Here, uncertainty refers to the distribution of dwells over the facial features. When entropy (as a measure for uncertainty) of the distribution of dwells is high, gaze is evenly distributed over the five AOIs. In this case, it is uncertain which AOI is fixated at a given moment. When entropy of the distribution of dwells across AOIs is zero, only one AOI is fixated. In

## 8. Gaze behavior during dyadic interaction

this case, it is certain which AOI is fixated at a given moment and as such the uncertainty in dwells over the AOIs is zero.

To confirm that entropy was indeed reflective of the amount of scanning by the confederate, a repeated-measures ANOVA on the confederate's entropy was run with confederate instructions (fixate eyes, scan face) as a within-subject factor and order (*fixate eyes instruction first*, *scan face instruction first*) as a between-subject factor. A significant main effect of confederate instructions ( $F(1,28) = 344.57, p < .001, \eta^2 = 0.93$ ) confirmed that entropy was lower when the confederate fixated the eyes than when he was instructed to scan the face. This is exactly as expected, as only fixating the eyes results in a more uneven distribution of dwells over the AOIs, and as such lower entropy. The main effect of order was non-significant ( $F < 1$ ), as was the interaction between confederate instructions and order ( $F < 1$ ), indicating that there were no differences in entropy between the *fixate eyes instruction first* and *scan face instruction first* order of instructions (as visible from the left panel of Figure 8.8).

To determine whether the amount of uncertainty (as measured by entropy) in the gaze behavior of the confederate elicited a change in the amount of uncertainty in the gaze behavior of the observers, a repeated-measures ANOVA on observers' entropy was run with confederate instructions (fixate eyes, scan face) as a within-subject factor and order (*fixate eyes instruction first*, *scan face instruction first*) as a between-subject factor. If uncertainty of the observers' gaze behavior was indeed dependent on the uncertainty of the confederate's gaze behavior, we would expect to see higher entropy for the observers when the confederate fixated the eyes of the observer. Again, we would expect the pattern we saw for the confederate in Figure 8.8 (left panel) to emerge for the observers. As visible from Figure 8.8 (right panel), this was not the case. The entropy of total dwell times to the AOIs was similar regardless of confederate instructions or the specific group participants were assigned to. A non-significant main effect of confederate instructions ( $F(1,28) = 1.55, p > .05$ ) revealed that the entropy of observers' gaze distribution did not differ based on the confed-

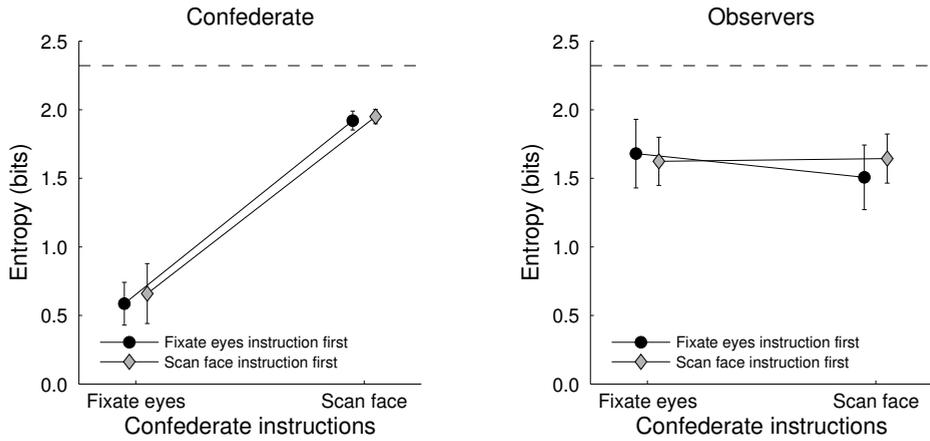


Figure 8.8.: Entropy of total dwell times for the confederate (left panel) and observers (right panel) as a function of the instructions given to the confederate in experiment 2. Separate lines depict the order in which the confederate fixated the eyes and scanned the face. A horizontal shift for one of the lines has been introduced on the x-axis to facilitate interpretation of the graph. The dashed line represents the maximum entropy for five AOIs at  $-\log_2(0.2)$ . Error bars depict 95% confidence interval.

## 8. Gaze behavior during dyadic interaction

erate's gaze behavior. Moreover, the order in which the confederate fixated the eyes and scanned the face of the observer did not affect the entropy of observers' gaze distribution ( $F < 1$ ). Finally, there was no interaction of order and confederate instructions ( $F(1,28) = 2.45, p > .05$ ). To sum, the distribution of the observers' gaze over the AOIs did not depend on the amount of time the confederate scanned the observer's face.

### 8.5.4. The effect of confederate gaze behavior – is it short lived?

A final alternative we aimed to exclude was that the effects of the confederate's gaze behavior on the gaze behavior of the observers are short lived and cannot be detected in an average over 150 seconds. We therefore investigated total dwell time for 10-second bins across the entire experiment (smaller bins yielded no extra information). As can be seen in the left panel of Figure 8.9, the confederate's total dwell time to the eyes increased sharply in the *fixate eyes instruction first* condition, and after 150 seconds, the slope decreased when the confederate started scanning the entire face region. This pattern was observed vice versa for the confederate in the *scan face instruction first* condition. As can be seen in the right panel of Figure 8.9, the lines for the observers' total dwell to the eyes in the *fixate eyes instruction first* and *scan face instruction first* condition overlap completely, refuting the alternative that the effects of the confederate's gaze behavior (i.e. looking straight into the eyes versus scanning the observers' face) are short lived. To sum, we find no evidence that the confederate's gaze behavior had any effect on the observers' gaze behavior.

## 8.6. General discussion

A plethora of studies have reported that humans have a bias for fixating the eye region in the faces of others. However, most studies have used static images or videos of faces, and the present study questioned whether this bias would generalize to gaze behavior to faces during social interaction. In order to answer this question, we designed a novel setup in which a duo could engage in social interaction while their eye movements

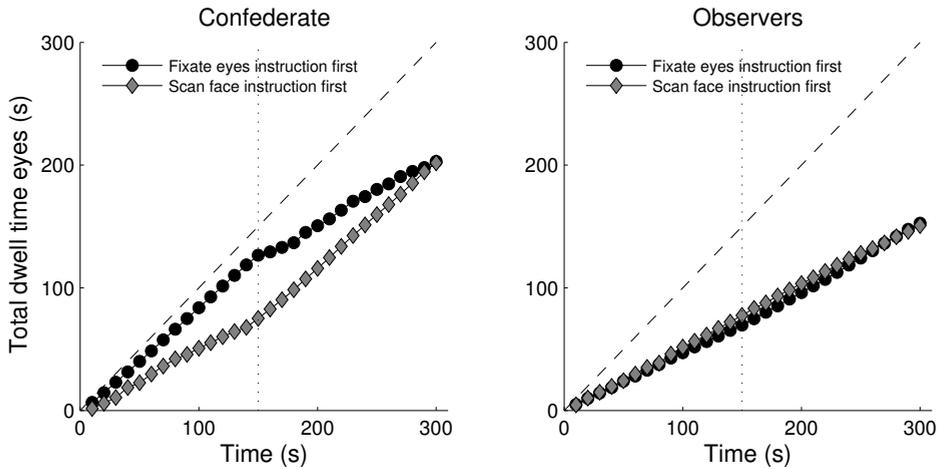


Figure 8.9.: Total dwell time to the eyes as a function of time in experiment 2 for the confederate (left panel) and the observers (right panel). As time in the experiment progressed, total dwell time to the eyes accumulated. Separate lines depict the *fixate eyes instruction first* and *scan face instruction first* order of instructions. For the confederate these lines depict clear differences, as his behavior was instructed to differ between the two orders of instructions. The dotted line represents the halfway-mark when the confederate switched from looking solely at the left eye to scanning the observers' face or vice versa. The dashed line represents the unity line.

## 8. Gaze behavior during dyadic interaction

were simultaneously recorded. A two-way video setup allowed simultaneous video and eye-movement recordings of both participants, without placing mobile eye-trackers on the participants' heads. Moreover, using a two-way video setup allowed us to track eye-movements on a 2-d plane, and perform semi-automatic AOI construction.

In experiment 1, we showed that there was a bias for fixating the eye region in the face of an interaction partner when that partner was physically present, corroborating the long-standing hypothesis that the eyes are an important source of information in faces. Moreover, we reported that the total amount of time spent looking at the eyes by either partner was highly correlated with the total amount of time the other partner spent looking at the eyes. Dwell durations in the eye region were twice as long as dwells to the nose or mouth region, indicating that when the eyes were looked at, they retained attention longer than the nose or mouth region. The average time that partners maintained eye contact – defined as the period of time in which both partners looked at each other's eyes – was less than half a second. This finding is a markedly lower estimate of the duration of eye contact than previously assumed. Earlier research estimates eye contact to take between 1-3 seconds (Argyle, 1972). Such estimates, however, were made in the absence of objective measurements of gaze behavior, or define eye contact to occur when to people look at each other's face, as opposed to a more restricted area in the present study. The shorter estimate of eye contact obtained here is intuitive, given that dwells were generally short and that gaze was directed at different facial features in quick succession. Finally, we examined the interplay of gaze behavior between the two partners more closely. Such an approach is similar to Ho et al. (2015), who investigated gaze as a means to start and end periods of talking. One example we report is that duos may either show a positive or inverse relation in gaze behavior as a function of time.

As total dwell times to the eyes were correlated between two partners in a duo in experiment 1, we proposed that partners might attune to the level of eye contact instigated by either partner. A change in the gaze behavior

of one partner should result in a change in the gaze behavior of the other partner. In experiment 2, a confederate posed as one of the partners in a duo, and followed instructions of where to look on the observer's face. The confederate either fixated the eyes of the observer, or scanned the entire face (i.e. by consecutively looking at instructed locations on the observer's face). Gaze behavior of the confederate did not affect the gaze behavior of the observers, thereby failing to corroborate our proposed model of attuning to the level of eye contact in social interaction.

Although the bias for fixating the eye region in faces has often been reported on (e.g. Birmingham, Bischof, & Kingstone, 2008a; 2008b; Henderson et al., 2005; Langton et al., 2000), even dating back to the work of Yarbus (1967), recent studies have questioned whether this bias may generalize to situations where there is actual interaction between people (Freeth et al., 2013; Laidlaw et al., 2011; Risko et al., 2012). Here we confirm that the bias to fixate the eye region generalizes to social interaction, at least for the case where people are required to observe each other (cf. Laidlaw et al., 2011). There are several caveats we should note, however. The bias for fixating the eye region does not necessarily mean that the eyes are the only relevant areas to fixate in a face. As has previously been proposed by Vö et al. (2012), humans may fixate the eyes, nose or the mouth depending on whether a person is talking or merely looking at the camera. In the present study, participants were asked to look towards each other, and no other instructions or restrictions on how to behave were given. While this was done in an attempt to elicit natural behavior and not focus participants on their own behavior, the question remains what interactions – apart from interactions in gaze behavior – actually occurred. In experiment 1, there was little to no talking between partners. In 7 of 9 duos, there was at least one period of laughter. In experiment 2, conversation occurred with 5 observers, and at least one period of laughter occurred in 17 out of 30 duos. While interaction in talking and laughing did occur in both experiments, it may have been limited and similar to the competitive duos rather than the cooperative duos in Jarick & Kingstone (2015). Given that there was limited talking, it is likely that the mouth was not a primary source of

## 8. Gaze behavior during dyadic interaction

information, and the eye region may have been most informative.

A second caveat to note is that participants observed each other through a two-way video setup. There was a physical barrier between the two participants and they only saw the face and shoulders of the other, not the entire body. Future research should examine whether this remaining degree of artificiality has a marked effect on gaze behavior to faces or not. A final caveat to note is that participants knew their eye movements were being tracked. Risko & Kingstone (2011) report that the belief that one's eye movements are being tracked affects his or her gaze behavior. In their study the amount of people looking at a provocative stimulus located at 90° head turn away was investigated. In the present study, however, gaze behavior to faces at a fine-grained level was investigated, and we deem it unlikely that large effects on gaze behavior as a result of the knowledge of being eye-tracked may have been present here. Given these caveats, we conclude that the way in which people observe the face of a partner whom is physically present does not appear to be fundamentally different to when the face is in a static image or video.

A feature critical to the experimental setup in the present study was that we were able to correlate gaze behavior of two partners with each other. We observed that the total amount of time spent looking at the eyes appeared to be duo-dependent. This finding is particularly relevant given two previous studies. Gobel et al. (2015) first reported that participants look shorter at the eyes of another person of high social rank when one believes that the other person may look back compared with when the other person won't look back. In addition, Jarick & Kingstone (2015) reported that the time two people can maintain eye contact depends on the relation between them (i.e. competitive versus cooperative). If the total amount of time people look each other in the eyes is indeed reflective of the degree to which their relationship is cooperative, competitive or based on differences in social rank, it may generally be the case that looking at the eyes is duo-dependent. However, when we explicitly manipulated the gaze behavior of one partner in the duo, this effect was not observed; not in the amount of time spent

looking at the eyes, not in the distribution of their gaze across the facial features, not even for a brief period of time. This might suggest that even if the amount of time partners look at each other's eyes is duo-dependent and reflective of their underlying relationship, gaze behavior is not the only cue to this underlying relationship. It may be that other non-verbal cues such as facial expressions are critical factors in predicting gaze behavior to a face during social interaction. Another possible explanation is that the explicit instructions given to the confederate were detrimental to the interaction itself: by substantially manipulating one partner in the duo, we might have eliminated the interplay of social cues between the two partners. However, we did still observe periods of talking and laughing for a number of duos, reflecting the fact that there was still interaction of some sort. Future research should critically examine whether these findings extrapolate to interaction of two people without a physical barrier or involved confederate. A final possibility is that the coupling of gaze behavior of two partners as a function of time may reveal more about the underlying relationship. We describe, for example, that gaze patterns appear to be direct or inversely related depending on the duo. Future research may focus on uncovering the precise dynamics of gaze behavior in social interactions. One approach would be to develop tools similar to those used by Ho et al. (2015). In their study, they used cross-correlational analysis to quantify the relation between gaze and periods of talking. Here, however, we would want to investigate the interplay between partners in gaze to four AOIs (eyes, nose, mouth, and non AOI). As the probabilities of looking at the four AOIs are dependent on each other, and the timing of the interplay may be duo-dependent, a cross-correlation analysis cannot capture this fully. The tools to analyze such models of interaction need first be developed.

The present study was conducted in the spirit of the Cognitive Ethology approach (Kingstone, 2009; Kingstone et al., 2008; Smilek et al., 2006), and by recording gaze behavior of two partners in a duo when they had the possibility to engage in interaction we were able to investigate how gaze behavior of the two partners was related. This method provides exciting possibilities for future research into the role of gaze behavior in social in-

## 8. Gaze behavior during dyadic interaction

teraction, of which we outline a few. In Autism Spectrum Disorder (ASD) research, several studies find reduced fixating of the eye region by individuals with ASD (Dalton et al., 2005; Hernandez, Metzger, Magné, & Bonnet-Brilhault, 2009; Klin, Jones, Schultz, Volkmar, & Cohen, 2002; Pelphrey et al., 2002), even early in development (Jones & Klin, 2013). However, several other studies do not find such differences between individuals with ASD and typically developing controls (Dapretto et al., 2005; McPartland, Webb, Keehn, & Dawson, 2010; Van der Geest, Kemner, Verbaten, & van Engeland, 2002). Speer, Cook, McMahon, & Clark (2007) investigated whether this might partly be because of social context, and report that individual with ASD only looked less at the eyes for complex social videos (i.e. 2 people in interaction), but not for static images of people or isolated people in video (see also Kemner & van Engeland, 2003, for an alternative explanation). By investigating gaze behavior to faces during social interaction we may learn more about social deficits in ASD, and how they may be influenced by interaction partners (see also De Jaegher, Di Paolo, & Gallagher, 2010). Moreover, given that gaze behavior to faces is more generally investigated in relation to psychopathology (e.g. in social phobia; Horley, Williams, Gonsalvez, & Gordon, 2003), future research may focus on how gaze behavior in social interaction may be predicted by (sub)clinical psychopathology. Such ideas are already gaining ground in the field of human-robot interaction (Broz, Lehmann, Nehaniv, & Dautenhahn, 2012; Damm et al., 2012).

To conclude, when two partners look at each other while there is the possibility of engaging in interaction, there is a bias to fixate the eyes. Moreover, the amount of time spent looking at the eyes appears duo-dependent, but not easily manipulated by changing the gaze behavior of one partner. Finally, it appears difficult to pinpoint exactly how this correlation of looking at the eyes between partners develops over time. We have outlined several important areas of application, which may benefit from the presented results.

## **Acknowledgements**

The authors would like to thank Gijs Holleman for valuable help in data collection. This work was supported by a Netherlands Organization for Scientific Research (NWO) VICI grant (No. 45307004) and by the Consortium on Individual Development (CID). CID is funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the NWO (Grant No. 024.001.003).

## References

- Argyle, M. (1972). *The psychology of interpersonal behavior*. Penguin Books, Harmondsworth, UK.
- Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The “reading the mind in the eyes” test revised version: A study with normal adults, and adults with asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry*, 42(2):241–251.
- Bindemann, M., Burton, A. M., Hooge, I. T. C., Jenkins, R., & de Haan, E. H. F. (2005). Faces retain attention. *Psychological Bulletin & Review*, 12(6):1048–1053.
- Birmingham, E., Bischof, W. F., & Kingstone, A. (2008a). Gaze selection in complex social scenes. *Visual Cognition*, 16(2-3):341–355.
- Birmingham, E., Bischof, W. F., & Kingstone, A. (2008b). Social attention and real-world scenes: The roles of action, competition and social content. *The Quarterly Journal of Experimental Psychology*, 61(7):986–998.
- Birmingham, E., Bischof, W. F., & Kingstone, A. (2009). Saliency does not account for fixations to eyes within social scenes. *Vision Research*, 49(24):2992–3000.
- Broz, F., Lehmann, H., Nehaniv, C. L., & Dautenhahn, K. (2012). Mutual gaze, personality, and familiarity: Dual eye-tracking during conversation. *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*.
- Dalton, K. M., Nacewicz, B. M., Johnstone, T., Schaefer, H. S., Gernsbacher, M. A., Goldsmith, H. H., Alexander, A. L., et al. (2005). Gaze fixation and the neural circuitry of face processing in autism. *Nature neuroscience*, 8(4):519–526.
- Damm, O., Malchus, K., Jaecks, P., Stenneken, P., Krach, S., Paulus, F., Becker, K., Jansen, A., Naber, M., Kamp-Becker, I., Einhaeuser-Treyer, W., & Wrede, B. (2013). Different gaze behavior in human-robot interaction in Asperger’s syndrome: an eye-tracking study. *2013 IEEE RO-MAN The 22nd IEEE International Symposium on Robot and Human Interactive Communication*.
- Dapretto, M., Davies, M. S., Pfeifer, J. H., Scott, A. A., Sigman, M., Bookheimer, S. Y., & Iacoboni, M. (2005). Understanding emotions in others: mirror neuron dysfunction in children with autism spectrum disorders. *Nature neuroscience*, 9(1):28–30.
- De Jaegher, H., Di Paolo, E., & Gallagher, S. (2010). Can social interaction constitute social cognition? *Trends in Cognitive Sciences*, 14(10):441–447.

- Dixon, N. F., & Spitz, L. (1980). The detection of auditory visual desynchrony. *Perception*, 9:719–721.
- Freeth, M., Foulsham, T., & Kingstone, A. (2013). What affects social attention? Social presence, eye contact and autistic traits. *PLOS One*, 8(1):e53286.
- Gallup, A. C., Chong, A., Kacelnik, A., Krebs, J. R., & Couzin, I. D. (2014). The influence of emotional facial expressions on gaze-following in grouped and solitary pedestrians. *Scientific Reports*, 4:5794.
- Gliga, T., Elsabbagh, M., Andravizou, A., & Johnson, M. (2009). Faces attract infants' attention in complex displays. *Infancy*, 14(5):550–562.
- Gobel, M. S., Kim, H. S., & Richardson, D. C. (2015). The dual function of social gaze. *Cognition*, 136:359–364.
- Goren, C. C., Sarty, M., & Wu, P. Y. K. (1975). Visual following and pattern discrimination of face-like stimuli by newborn infants. *Pediatrics*, 56(4):544–549.
- Henderson, J. M., Williams, C. C., & Falk, R. J. (2005). Eye movements are functional during face learning. *Memory & Cognition*, 33(1):98–106.
- Hernandez, N., Metzger, A., Magné, R., & Bonnet-Brilhault, F. (2009). Exploration of core features of a human face by healthy and autistic adults analyzed by visual scanning. *Neuropsychologia*, 47:1004–1012.
- Hessels, R. S., Kemner, C., van den Boomen, C., & Hooge, I. T. C. (2016). The area-of-interest problem in eyetracking research: A noise-robust solution for face and sparse stimuli. *Behavior Research Methods*, 48(4):1694–1712.
- Ho, S., Foulsham, T., & Kingstone, A. (2015). Speaking and listening with the eyes: Gaze signaling during dyadic interactions. *PLOS One*, 10(8):e0136905.
- Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press.
- Hooge, I., & Camps, G. (2013). Scan path entropy and arrow plots: Capturing scanning behavior of multiple observers. *Frontiers in Psychology*, 4:996.
- Horley, K., Williams, L. M., Gonsalvez, C., & Gordon, E. (2003). Social phobics do not see eye to eye: A visual scanpath study of emotional expression processing. *Journal of anxiety disorders*, 17:33–44.
- Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40:1489–1506.
- Jarick, M., & Kingstone, A. (2015). The duality of gaze: Eyes extract and signal social information during sustained cooperative and competitive dyadic gaze. *Frontiers in Psychology*, 6:1423.

## 8. Gaze behavior during dyadic interaction

- Johnson, M. H., Dziurawiec, S., Ellis, H., & Morton, J. (1991). Newborns' preferential tracking of face-like stimuli and its subsequent decline. *Cognition*, 40:1–19.
- Jones, W., & Klin, A. (2013). Attention to eyes is present but in decline in 2–6-month-old infants later diagnosed with autism. *Nature*, 504:427–431.
- Kemner, C., & van Engeland, H. (2003). Autism and visual fixation. *American Journal of Psychiatry*, 160(7):1358–1359.
- Kingstone, A. (2009). Taking a real look at social attention. *Current Opinion in Neurobiology*, 19:52–56.
- Kingstone, A., Smilek, D., & Eastwood, J. D. (2008). Cognitive ethology: A new approach for studying human cognition. *British Journal of Psychology*, 99(3):317–340.
- Klin, A., Jones, W., Schultz, R., Volkmar, F., & Cohen, D. (2002). Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Archives of General Psychiatry*, 59:809–816.
- Laidlaw, K. E. W., Foulsham, T., Kuhn, G., & Kingstone, A. (2011). Potential social interactions are important to social attention. *Proceedings of the National Academy of Sciences*, 108(14):5548–5553.
- Langton, S. R. H., Law, A. S., Burton, A. M., & Schweinberger, S. R. (2008). Attention capture by faces. *Cognition*, 107:330–342.
- Langton, S. R. H., Watt, R. J., & Bruce, V. (2000). Do the eyes have it? Cues to the direction of social attention. *Trends in Cognitive Sciences*, 4(2):50–59.
- McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1):30–46.
- McPartland, J. C., Webb, S. J., Keehn, B., & Dawson, G. (2010). Patterns of visual attention to faces and objects in autism spectrum disorder. *Journal of Autism and Developmental Disorders*, 41(2):148–157.
- Pelphrey, K. A., Sasson, N. J., Reznick, J. S., Paul, G., Goldman, B. D., & Piven, J. (2002). Visual scanning of faces in autism. *Journal of Autism and Developmental Disorders*, 32(4):249–261.
- Risko, E. F., & Kingstone, A. (2011). Eyes wide shut: Implied social presence, eye tracking and attention. *Attention, Perception & Psychophysics*, 73(2):291–296.
- Risko, E. F., Laidlaw, K. E. W., Freeth, M., Foulsham, T., & Kingstone, A. (2012). Social attention with real versus reel stimuli: Toward an empirical approach to concerns about ecological validity. *Frontiers in Human Neuroscience*, 6(1):143.

- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423.
- Smilek, D., Birmingham, E., Cameron, D., Bischof, W., & Kingstone, A. (2006). Cognitive ethology and exploring attention in real-world scenes. *Brain Research*, 1080:101–119.
- Speer, L. L., Cook, A. E., McMahon, W. M., & Clark, E. (2007). Face processing in children with autism: Effects of stimulus contents and type. *Autism*, 11(3):265–277.
- Van der Geest, J. N., Kemner, C., Verbaten, M. N., & van Engeland, H. (2002). Gaze behavior of children with pervasive developmental disorder toward human faces: A fixation time study. *Journal of Child Psychology and Psychiatry*, 43(5):669–678.
- Võ, M. L. H., Smith, T. J., Mital, P. K., & Henderson, J. M. (2012). Do the eyes really have it? Dynamic allocation of attention when viewing moving faces. *Journal of Vision*, 12(13):3.
- Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *Journal of Strength and Conditioning Research*, 19(1):231–240.
- Yarbus, A. L. (1967). *Eye movements and vision*. Plenum Press, New York.

## 8. Gaze behavior during dyadic interaction



Figure 8.10.: The experimental setup used in the study.

## 8.7. Appendix

### 8.7.1. Eye-tracking setup

Figure 8.10 depicts the experimental setup used in the present study. Participants were positioned on either end of the table on an office chair, in front of one of the wooden boxes. The experimenter was positioned in front of the three screens running the eye-tracking and video-recording software. Figure 8.11 depicts one end of the two-way video setup. The participant looks at a half-silvered mirror, which reflects the screen lying on the bottom of the wooden box. Just below the half-silvered mirror is the eye tracker. The black cloth covering the lower half of the wooden box prevents the participant from looking directly at the screen instead of at the half-silvered mirror. A video example of the data recorded from this setup is available at <http://www.royhessels.nl/research.html>.

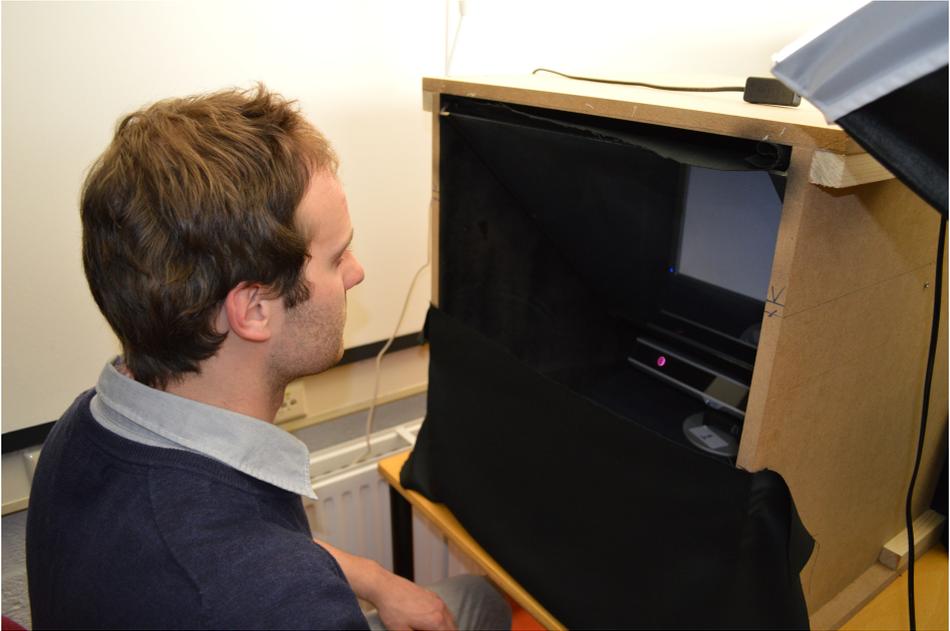


Figure 8.11.: A participant in front of one end of the two-way video setup.

### 8.7.2. Scarf plots

Figure 8.12 depicts scarf plots for the four duos not included in the Results section of experiment 1. Figure 8.13 depicts scarf plots for the confederate in experiment 2, separated for the two possible orders of instruction: *fixate eyes instruction first*, and *scan face instruction first*. Note that the confederate was instructed to look at 10 different locations on the face (see Methods section of experiment 2) of the observers, while the confederate's gaze is here represented into four AOIs. Moreover, gaze behavior may be shifted slightly for each observer because of differences in the precise moment of instruction onset.

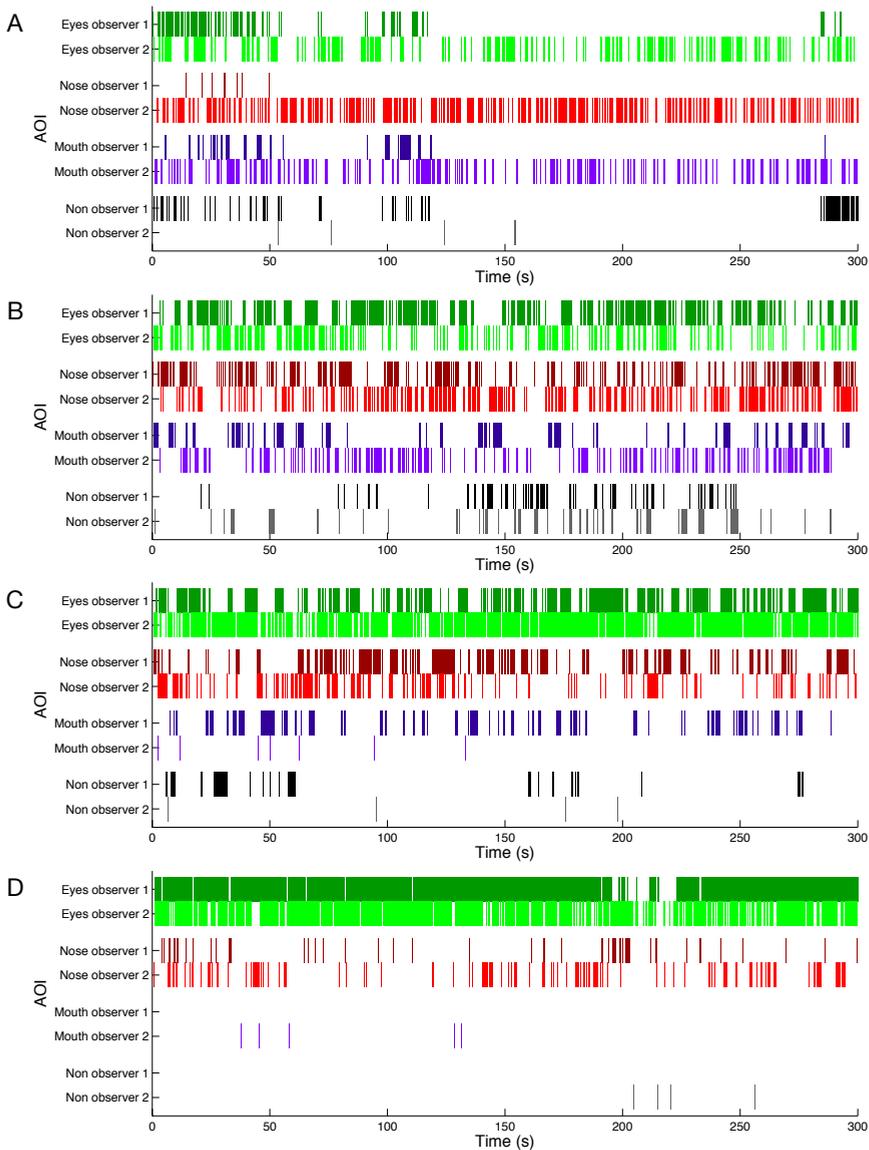


Figure 8.12.: Scarf plots of dwells to the four AOIs for 4 duos (panels A-D) in order of the total dwell time to the eyes in experiment 1. Each filled rectangle represents one dwell with its length equal to the duration of that dwell. The dark colored bars in each scarf plot belong to observer 1, whereas the lighter colored bars belong to observer 2. Bars are grouped by each AOI (eyes, nose, mouth, and non AOI). This allows easy comparison of dwells to the AOIs as a function of time between participants.

## 8. Gaze behavior during dyadic interaction

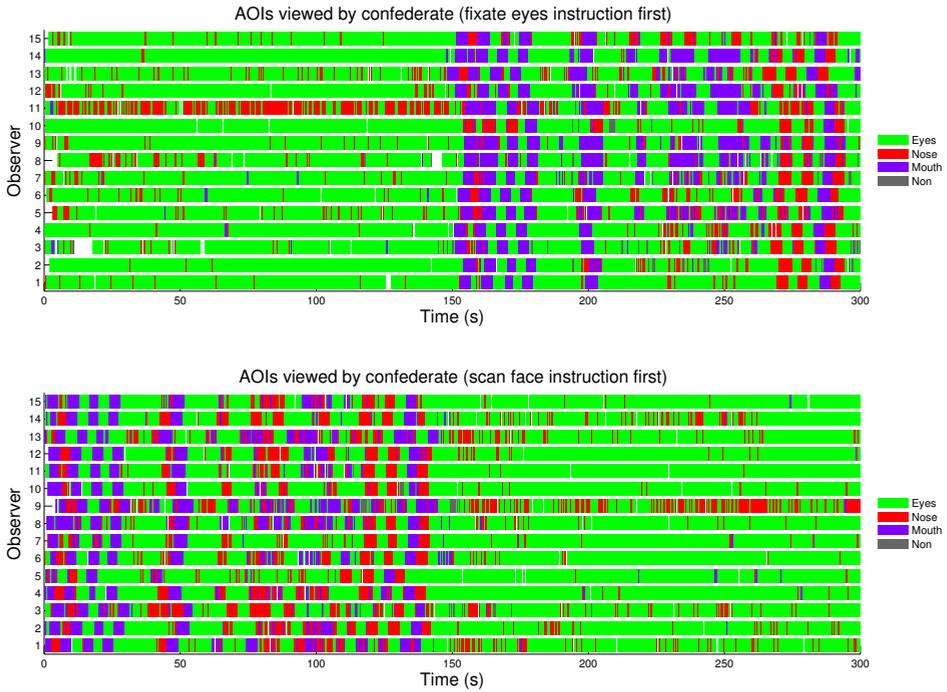


Figure 8.13.: Scarf plots of dwells to the four AOIs for the confederate in experiment 2. Each filled rectangle represents one dwell with its length equal to the duration of that dwell. Each row represents the gaze behavior of the confederate on the face of one observer. The top panel depicts the observers in the fixate eyes instruction first, and the bottom panel depicts the observers in the scan face instruction first order of instructions.

## **9. Eye contact takes two – capturing gaze behavior of subclinical autism and social anxiety in dyadic interaction**

Paper submitted as:

Hessels, R. S., Holleman, G. A., Cornelissen, T. H. W., Hooge, I. T. C., & Kemner, C. (2017). Eye contact takes two – capturing gaze behavior of subclinical autism and social anxiety in dyadic interaction.

Author contributions:

RH, GH, TC, IH, CK designed the study. RH, GH collected the data. RH, GH analyzed the data. RH, GH, TC, IH, CK interpreted the data. RH drafted the paper. RH, GH, TC, IH, CK finalized the paper.

## **Abstract**

Over the last decades, researchers have investigated face processing from many different perspectives, and while this has resulted in a plethora of knowledge on the sub-systems of facial information processing, recent evidence suggests that generalizability of these findings to social settings may be limited. The main argument is that in social interaction, the content of faces is more dynamic and dependent on the interplay between interaction partners, than the content of a non-responsive face as portrayed in a typical experiment. The question beckons whether gaze behavior to faces as investigated with non-responsive stimuli generalizes to faces in interaction. In the present study we investigated gaze behavior in social interaction using a novel dual eye-tracking setup capable of recording gaze with high resolution during interaction. More specifically, we investigated whether gaze behavior in interaction is related to (sub)-clinical traits of Autism Spectrum Disorder (ASD), and Social Anxiety Disorder (SAD). We report that gaze behavior of individuals scoring high on ASD and SAD to the face of an interaction partner corroborates long-standing findings obtained using static pictures and videos. Moreover, we report that pairs of observers scoring high on ASD traits were engaged for a shorter time in instances when both partners looked at each other's eyes, but engaged for a longer time in instances when only one partner looked at the eyes of the other, compared to pairs of observers scoring low on ASD traits. Additionally, pairs of observers scoring high on SAD traits were more often engaged in one-way eye gaze (only one partner looks at the eyes of the other), yet for shorter duration than pairs of observers scoring low on SAD traits. These findings provide intriguing possibilities for the investigation of gaze behavior in interaction, and attest to the sensitivity of gaze behavior in dyadic interaction to (sub)clinical psychopathology.

Faces carry information that is crucial to social interaction. They convey information on one's identity and emotional expression, for example, which are important for our willingness to engage in, or respond to bids for, social interaction. Over the last decades, researchers have investigated face processing by compartmentalizing face processing into smaller problems. For example, researchers have investigated how gaze direction is inferred from one's eyes and head orientation (e.g. Langton et al., 2000; Langton, 2000), how shifts of attention occur based on that gaze direction (Frischen et al., 2007), how emotion from a face is perceived or processed (Leppänen & Nelson, 2009), how one's mental state may be conveyed in the eye region (Baron-Cohen et al., 2001a), in what manner information is retrieved when viewing faces (e.g. Yarbus, 1967; Võ et al., 2012), which information is favored when learning faces (Henderson et al., 2005), and many more. This has for the greater part been done by manipulating specific aspects of static faces. While this compartmentalization has resulted in a wealth of knowledge on the subsystems of face processing, this approach has recently been questioned by several researchers for its generalizability to social settings, where the content of faces is more holistic, dynamic and dependent on the interplay between interaction partners (Smilek et al., 2006; Kingstone et al., 2008; Kingstone, 2009; Risko et al., 2012).

The concern for the lack of generalizability of studies on face processing to social settings is based on several arguments. First, the context of a social setting appears to play an important role in face processing, and particularly in gaze behavior to faces. Birmingham et al. (2008) for example, showed that increasing the number of people in a static scene, combined with the activity portrayed (i.e. playing a game versus inactivity) increased the relative time spent looking at the eyes. A second argument is that the presence of an actual person as opposed to a representation (e.g. a photograph, video, or animated model) modulates gaze behavior. Indeed, Laidlaw et al. (2011) report that participants looked for a longer total duration at the videotape of a person than at a person actually being present in the same room. Gobel et al. (2015) furthermore showed that the belief that one's eye movements and face are being recorded, and will

## 9. Gaze in interaction in subclinical ASD and SAD

be shown to another participant, modulated the ratio of time spent looking at the eyes and time spent looking at the mouth (even based on the social rank of the person looked at). Third, the eyes of the observer are not only important for information uptake, but may provide valuable signals to others. Research has therefore examined whether the gaze behavior of one person affects that of the other. Freeth et al. (2013), for example, report that participants looked more at the face of the interviewer when she made eye contact with participants, compared to when the interviewer did not engage in eye contact. These studies emphasize that the presence of another person, as well as the behavior of that person, matters for gaze behavior. In a recent review, Risko et al. (2012) state that merely investigating passive viewing of representations of others may overlook important dynamics of gaze behavior to faces, compared to situations in which social partners are physically present.

The insight that gaze behavior in social interaction is likely to be quite different from gaze behavior to non-interactive representations of others (pictures, videos, etc.) is not only important for face processing in general, but also bears on research in psychopathology where social impairments are fundamental to certain disorders, e.g. Autism Spectrum Disorder (ASD) and Social Anxiety Disorder (SAD). Social impairments in these disorders have often been investigated by using eye-tracking technology to gain insights into gaze behavior to representations of others (e.g. Guillon et al., 2014). Here, looking at the eyes of the representation of a person is considered as ‘eye contact’ and is studied as a model for social interaction (Senju & Johnson, 2009a,b). Given that a plethora of knowledge on these disorders derives from research using such simple representations of persons, the question beckons whether this generalizes to gaze behavior in social interaction. Although the advance of wearable eye-tracking technology (often in the form of glasses called ‘mobile eye trackers’) has made it possible to investigate gaze behavior between interacting people, there are several drawbacks. First, when these mobile eye trackers are used to study gaze behavior in interaction, the resolution of the eye-trackers is limited to whether one looks at a face or not, and precludes more detailed information of gaze

on facial features (e.g. Gullberg & Holmqvist, 2006; Broz et al., 2012; Damm et al., 2013). While the resolution to analyze such detailed gaze information is getting better with new eye-tracking models, a second drawback may still limit its use for fine-grained analysis, namely that analyzing eye-movement data from mobile eye-trackers is often a time-intensive manual coding process. Thus the trade-off between a controlled, high-resolution laboratory study with high quality eye-tracking data and automatic analysis and a lower-resolution mobile eye-tracking study in which people can interact (with poorer eye-tracking data quality and often manual analysis) seems a difficult one to break. Here we present one attempt to do so by investigating high-resolution gaze behavior to faces in dyadic interaction using a dual eye-tracking setup, and link gaze behavior to subclinical ASD and SAD traits. First, research on gaze behavior to faces in ASD and SAD is briefly reviewed, after which we introduce the specific research questions.

## 9.1. Gaze behavior to faces in Autism Spectrum Disorder

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterized by “*persistent deficits in social communication and social interaction*”, and “*restricted, repetitive patterns of behavior, interests, or activities*” in the Diagnostic and Statistical Manual of Mental Disorders 5 (DSM-5; American Psychiatric Association, 2013). One typical example of the deficits in social communication and social interaction is abnormal eye contact. The first study to use eye-tracking methodology to investigate how exactly individuals with ASD scan faces reported that adults with ASD looked less at the core features (eyes, nose and mouth) in photographs of faces compared to typically-developing (TD) adults (Pelphrey et al., 2002). Specifically, the adults with ASD spent less time looking at the eyes compared to TD adults. This finding of reduced time spent looking at eye region as a model for abnormality in making eye contact (and perhaps social interaction as a generalization) seems promising, particularly if the amount of time spent looking at the eyes may serve as a

## 9. Gaze in interaction in subclinical ASD and SAD

measurable index of impairment in social interaction. However, subsequent research has proven inconsistent in reporting reduced looking time to the eyes in ASD. Several studies report that adults (Sterling et al., 2008; Hernandez et al., 2009), adolescents (Klin et al., 2002; Dalton et al., 2005), school-aged children (Rice et al., 2012) and toddlers (Jones et al., 2008) with ASD spend less time looking at the eyes compared to TD individuals. Moreover, 6-month-old infants at risk for ASD looked less at the eyes of their mother when that mother maintained a still face halfway through a period of normal interaction compared to infants who were not at risk for ASD (Merin et al., 2007). Jones & Klin (2013), furthermore report that infants at-risk for ASD who later on received a diagnosis of ASD do not look less at the eyes compared to TD infants in the first 6 months after birth, but show diminished looking at the eyes hereafter. On the other hand, there are also reports that adults (Rutherford & Towns, 2008) and children (Van der Geest et al., 2002b; Dapretto et al., 2006; McPartland et al., 2011) with ASD do not look less at the eyes compared to TD individuals. Chawarska & Shic (2009) furthermore report that 4-year-olds but not 2-year olds with ASD look less at the core features of a face, but no differences on time spent looking at the eyes are reported. Concluding, the evidence for reduced looking time to the eyes in ASD is, in the least, inconsistent.

One explanation for the inconsistencies in gaze behavior to faces in ASD was posited by Speer et al. (2007). They suggest that the social context of the stimuli used may be very important. In their study photographs and videos of either one or more persons were employed. Individuals with ASD differed only from TD individuals in the social dynamic (videos of more people) condition, in which reduced looking time to the eyes was observed. While the social context may appear to reconcile differences in looking time to the eyes across studies, this does not reconcile differences in increased looking time to the mouth (Klin et al., 2002; Rice et al., 2012). However, one may question how ‘social’ the context was even in the condition where videos of more people were used. As a recent review on this topic concludes that *“the available evidence at present suggests that the re-*

*duced fixation on the eyes in ASD is most prominent under conditions of high cognitive demand*” (p. 1211), which may refer to “... *complex and cognitively demanding face stimuli ...*” (Senju & Johnson, 2009a, p. 1208–1209), investigating gaze behavior to faces in actual interaction may be a promising perspective for ASD research. Most importantly, using high-resolution eye-tracking techniques with automatic analyses in interaction are required, as inconsistencies in gaze behavior to faces in ASD were already reported when subjective observation techniques (e.g. video coding) were the prevailing methodology (Van der Geest et al., 2002a).

## 9.2. Gaze behavior to faces in Social Anxiety Disorder

Social Anxiety Disorder (SAD) is characterized in the DSM-5 as marked fear or anxiety about social situations where one is under possible scrutiny by others. Importantly, the fear or anxiety is out of proportion to the threat posed by the social situation. Moreover, individuals with SAD may show “*inadequate eye contact*” in social situations (American Psychiatric Association, 2013). Eye-tracking studies on gaze behavior to faces report that individuals with SAD look less at the facial features (eyes, nose, mouth), and particularly the eyes, compared to TD adults (Horley et al., 2003; Moukheiber et al., 2010). More specifically, Moukheiber et al. (2010) report that individuals with SAD look less often, and for shorter durations at the eyes, whereas Horley et al. (2004) report less looking at the eyes only for angry faces. A study investigating gaze behavior to faces in a general population report that individuals scoring high on social phobia traits show no avoidance of direct gaze (i.e. less looking at the eyes of a person in a video looking straight ahead), but are more aroused, as evidenced by increased heart rate compared to individuals scoring low on social phobia traits (Wieser et al., 2009).

Although these studies show altered gaze behavior to faces in SAD, particularly with regard to the eye region, the SAD research field also calls for “*more naturalistic social experimental tasks*” (Horley et al., 2004, p. 52). Wieser et al. (2009), for example, claim that a lot of social phobia research

## 9. Gaze in interaction in subclinical ASD and SAD

comes from “*real interactions, which do not allow valid and reliable assessment of dependent variables*” (p. 94). Here again, a reference is made to observation studies, which do not offer the objectivity of eye-tracking studies. While Wieser et al. (2009) tackle the problem of the representativeness of their experiment for a social setting by using video clips as an ideal middle-ground between the control of a laboratory and ‘real interactions’, the ideal situation would be a merger of the two.

### 9.3. The present study

In the present study we investigate gaze behavior in social setting, by concurrently recording eye movements of two people in dyadic interaction. We do so by using a novel social interaction dual eye-tracking setup that allows for the registration of high-resolution gaze behavior of two observers looking at each other’s faces. Importantly, this setup allows for largely automatic data analyses. Using this setup, we have already reported that the typical biases of gaze to faces as investigated using static faces are present in dyadic interaction as well (Hessels et al., 2017). Specifically we investigate the relation between ASD and SAD traits in a general population and gaze behavior in interaction. This is highly relevant given the tendency of the last decades of compartmentalizing face processing in sub problems, and the inconsistencies on gaze behavior to faces in ASD and SAD. Investigating gaze behavior in interaction in relation to ASD and SAD might prove very valuable in assessing the sensitivity of gaze in interaction to (sub)clinical psychopathology, as differences between clinical groups and control are not always clear.

By concurrently recording eye movements of two participants we are able to investigate gaze behavior one-way – where does one look on the face of an other and how is this related to (sub)clinical psychopathology? Second, it allows us to investigate gaze behavior two-way – how is gaze behavior in interaction related to (sub)clinical psychopathology? Specifically, it allows us to investigate paired gaze states – the combination of gaze of two participants, for example ‘eye contact’ or ‘gaze aversion’. Regarding the one-way

Table 9.1.: Descriptive statistics of the participants.

	All participants	Individual data set (one-way)	Paired data set (two-way)
Sample size	96	65	42
Age	24.52 ( $sd = 3.91$ )	25.00 ( $sd = 4.31$ )	25.51 ( $sd = 5.05$ )
Gender	46 male, 50 female	37 male, 28 female	23 male, 19 female
Pair configuration	18 male, 22 female, 56 mixed	14 male, 14 female, 37 mixed	12 male, 8 female, 22 mixed
Non-Dutch speakers	4	3	2

analysis, we hypothesize scores on ASD and SAD questionnaires to be negatively correlated with the time spent looking at the eyes. Participants scoring higher on either trait look less at the eyes compared to participants scoring lower on that trait. Given the inconsistencies in previous research we will investigate whether ASD and SAD traits are positively correlated with time spent looking at other facial areas (nose and mouth). Regarding the two-way analysis, we will explore the relation between ASD and SAD traits and the total time, duration and frequency of paired gaze states.

## 9.4. Methods

### 9.4.1. Participants

One-hundred participants were recruited amongst the students, employees and visitors at the Faculty of Social and Behavioral Sciences of Utrecht University. All participants gave written informed consent prior to the experiment. Data of 4 participants were lost due to either an incorrect eye-tracking protocol ( $n = 2$ ) or because data were accidentally overwritten ( $n = 2$ ). Participants were arranged in pairs for the experiment. Descriptive statistics of the remaining 96 participants, as well as those of the groups after exclusion based on data quality (see section *Data quality*), are given in Table 9.1.

### 9.4.2. Apparatus & Stimuli

We wanted to map gaze positions of two observers onto the face of the other. We chose to do this using a Skype-like two-way video setup, in which a live feed from one observer was presented to the other, and vice versa. However, a Skype setup features the problem that a person does not appear to look one in the eye. The main reason for this is that the camera is placed above the screen. One can only appear to look another in the eye by looking straight into the camera. However, when looking in the camera, one cannot look at the image of the other at the same time. To circumvent this problem, we designed a setup similar to an autocue system TV presenters use, which allows a person to look straight into the camera and onto the image at the same time. This was done by placing a camera behind a half-silvered mirror reflecting the image of the other observer (see Figure 9.1).

The social interaction dual eye-tracking setup used in this study, introduced in Hessels et al. (2017), is depicted in Figure 9.1. The setup consists of two wooden boxes placed at either end of an office table. Inside of each wooden box was a half-silvered mirror, which reflected the screen lying in the bottom of the box. Behind the half-silvered mirror a webcam was placed. The webcam could record video of the participant through the mirror. This video image was subsequently presented on the screen of the other participant. Two Logitech webcams recorded video frames of 800 by 600 pixels at 30 Hz, and were presented at a resolution of 1024 by 768 pixels in the center of the 1680 by 1050 pixels available on the screen. The video frames were surrounded by a black border. The live-feeds were concurrently presented to the other participant and recorded to disk. Two SMI RED eye-tracking systems running at 120 Hz were used for the collection of eye-tracking data during the presentation of the live-feed of the observers. A parallel port connection was used to communicate to the eye-tracker computers that the video stream began and ended. Further technical details are provided in Hessels et al. (2017). Participants could see each other only through the dual eye-tracking setup as if they were positioned at a distance

of 1.36 m. Participants could hear each other, as they were both located in the same room.

### **9.4.3. Questionnaires**

ASD and SAD traits were assessed using short questionnaires. ASD was assessed using the Autism-spectrum quotient (AQ) developed by Baron-Cohen et al. (2001b), using the Dutch translation by Hoekstra et al. (2008). The AQ was developed to be a brief, self-administered instrument capable of assessing autism-spectrum traits in adults with normal intelligence. The AQ contains 50 questions, 10 of which each assess a different area: social skill, attention switching, attention to detail, communication, and imagination. AQ scores were calculated as described in Baron-Cohen et al. (2001b), range between 0 and 50, and higher values indicate more ‘autistic-like’ behavior associated with that particular area.

SAD was assessed using the Dutch social anxiety scale (SAS) developed by Willems et al. (1973). The SAS was developed to measure social anxiety as a disposition, and in particular the fear or anxiety of negative evaluation in social situations. The 24 questions tap into situations where the person stands out, situation in which evaluation or judgement of a person may occur, new and unexpected situations, and informal contact situations. Scores may vary between 0 (not socially anxious at all) to 96 (extremely socially anxious).

Given the international character of the University setting, a few international participants were recruited, and English translations of the questions were used for these participants. Depending on the specific analysis, this number was 4 or less (see Table 9.1).

### **9.4.4. Procedure**

Upon entering the lab, participants were placed at a table with a screen in between them to privately read and sign informed consent forms. Hereafter, participants entered the experimental room and were seated at each

9. Gaze in interaction in subclinical ASD and SAD

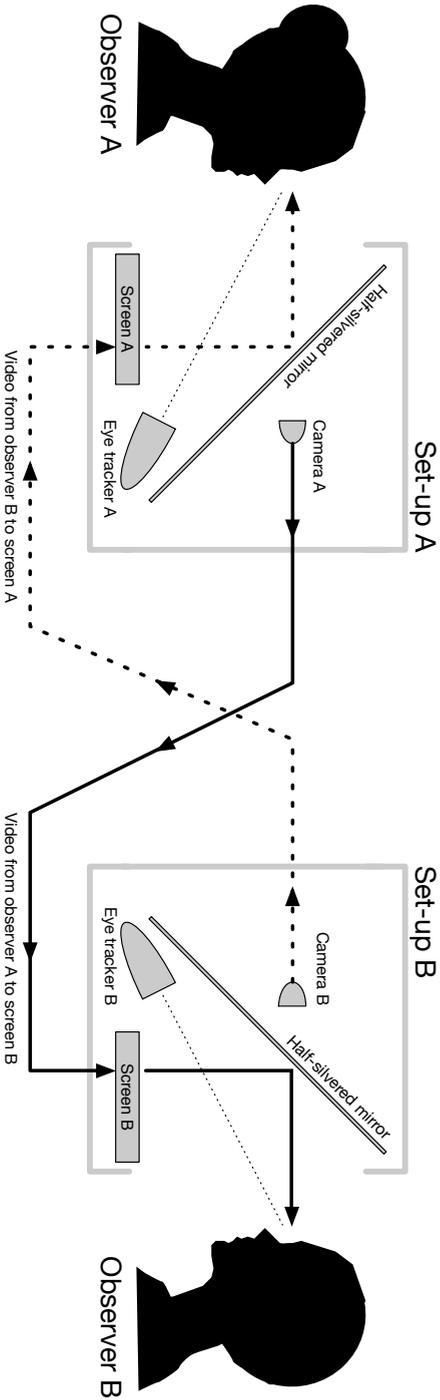


Figure 9.1.: Schematic side view of the dual eye-tracking setup. Observer A is recorded through a half-silvered mirror by Camera A. The solid line represents the flow of the video image of Observer A to Screen B, and is subsequently reflected of the half-silvered mirror to Observer B. As the eyes of Observer A are depicted on Screen B at the same height as Camera B is recording (and vice versa for Observer B to Observer A), both observers appear to look straight ahead when they look at the eyes of the other. Both sides of the setup were placed in the same room, with the observers at either side.

end of the dual eye-tracking setup. Using a height-adjustable chair, participants were placed such that their eyes were at the same height as the webcam filming them and the distance of their eyes to the eye tracker was approximately 70 cm. By placing participants with their eyes at webcam height, we ensured that when they looked straight ahead they would look directly into the camera. Moreover, this ensured that participants would also look directly on the eyes of the image of the other person portrayed on the half-silvered mirror. After positioning, a 5-point calibration procedure was run, succeeded by a 4-point validation procedure.

After validation, participants were instructed to ‘look at each other’ for a total of 5 minutes, identical to Hessels et al. (2017). No specific instructions were given to maintain eye contact, nor were there any restrictions on talking. The instructions were kept to a minimum to allow participants to behave as natural as possible in spite of the experimental setting. After the instruction the live-video stream and data recording were started. After 5 minutes the experimenter manually ended the experiment. Following the experiment participants were again seated at the table with the screen in between participants. Participants were then asked to fill in the self-report questionnaires through LimeSurvey, an online-survey tool.

#### **9.4.5. Data analysis**

In order to map gaze coordinates of both observers to the video recorded from the other observers, eye-tracking data were analysed as outlined in Hessels et al. (2017). In short, eye-tracking data were trimmed to the start and end of the videos, and processed using custom software written in MATLAB R2014b. Area of Interest (AOI) centers were manually determined for the left eye, right eye, nose, and mouth. Hereafter, face-tracking software automatically tracked the location and orientation of the participant’s face through the video and updated the location of the eyes, nose, and mouth for each video frame. As the video was recorded at 30 Hz, and eye-tracking data needed to be mapped to this video, eye-tracking data were down sampled using a moving-window average. For each video frame,

## 9. Gaze in interaction in subclinical ASD and SAD

the corresponding down-sampled gaze coordinate of the participant was assigned to one of four AOIs – left eye, right eye, nose, and mouth – using the Limited-Radius Voronoi Tessellation (LRVT) method (Hessels et al., 2016). The LRVT method assigns a gaze coordinate to the nearest AOI provided that its distance does not exceed the limited radius. Previous research has shown that large AOIs implemented using the LRVT method using large radii are most robust to noise in sparse stimuli such as faces (Hessels et al., 2016). The LRVT-radius was set to  $4.0^\circ$ . To compare, the average distance between the left eye and right eye AOI center across all videos was  $3.0^\circ$ , and the average distance from each AOI to its closest neighbor (or AOI span) was  $1.9^\circ$ <sup>1</sup>.

After gaze was assigned to the AOIs by the analysis software, dwells were computed in order to determine how long and how often observers looked at the AOIs. Dwells on the same AOI interspersed with a dwell of one video frame (e.g. a 33 ms dwell on the nose between two dwells on the left eye) were merged into one dwell. As fixations are generally longer than 100 ms, dwells shorter than 120 ms were excluded. Total dwell times (Holmqvist et al., 2011, p. 389) to the left eye, right eye, nose, and mouth were calculated by summing the durations of all dwells to the corresponding AOI. The total dwell times to the left and right eye were also summed to create a combined AOI for the eyes. A non-AOI was used for all frames in which gaze was available, but not on any of the three other AOIs; this included all gaze data on or off the screen, but not lost data.

For the two-way gaze analysis, three gaze states were defined. Two-way eye gaze refers to the situation where both participants look at the eyes AOI. One-way eye gaze refers to the situation where one participant looks at the eyes AOI, but the other does not. No eye gaze refers to the situation where both participants look somewhere else (e.g. nose, mouth, or non-AOI). While two-way gaze is sometimes referred to as ‘eye contact’, and no eye gaze as ‘averted gaze’, we refrain from using these terms, as it may

---

<sup>1</sup>These angles are reported under the assumption that participants remained at 81 cm from the screen and in the center of the camera image

imply something special or intentional about these gaze states. For all three gaze states, the frequency of occurrences, the mean duration of an occurrence, and the total time of the occurrences were calculated.

## 9.5. Results

### 9.5.1. Eye-tracking data quality

Validation error after calibration was approximately  $0.7^\circ$ . Upon inspection of the raw eye-tracking data, two problems were identified in some of the observers. First, it appeared that there was a large proportion of data loss for a number of observers (i.e. no gaze coordinate was reported). Second, for some observers the gaze coordinate appeared to be highly unstable, and change position very rapidly. In order to quantify these problems, two measures were calculated. First, the amount of time without dwells on any of the AOIs was calculated by subtracting the total dwell times on the left eye, right eye, nose, mouth and non-AOI from 300 s (the duration of the experiment). The measure contains not only periods of data loss (due to blinks or tracking problems), but also dwells shorter than the minimum allowed dwell time. Second, the root mean square (RMS) displacement was computed as an estimate for the rapid position change observed in the raw data. This RMS displacement was calculated as the root of the average squared position change from one video frame to the next. Note that this RMS displacement contains all eye movements and cannot therefore be compared with RMS noise as an estimate for the precision of eye-tracking data, which is generally calculated during fixation (Holmqvist et al., 2011, p. 360–362). Each participant is represented as one point in Figure 9.2 based on these two measures. A cluster of participants can be seen in the bottom left corner. These participants had short times without dwells on any AOI and small RMS displacement, whereas the other participants either scored high on RMS displacement or time with dwells. The latter participants were therefore excluded in the analysis of the individual gaze behavior. The numerical criteria corresponding to the selection of excluded participants were 1) over 70 s without dwells on any AOI or 2) RMS dis-

## 9. Gaze in interaction in subclinical ASD and SAD

placement above 3°. For the analysis on paired gaze states, only pairs with both participants in the individual analysis were selected. Descriptives of the individual and paired data sets are given in Table 9.1.

### 9.5.2. Individual gaze data

The individual gaze data were used to answer the question where people look on the face of another and how this is modulated by (sub)clinical psychopathology. We hypothesized scores on ASD and SAD questionnaires to be negatively correlated with the total dwell time at the eyes. As the duration of the experiment is fixed, a decrease in total dwell time at the eyes should be accompanied by an increase in total dwell time at another location. Therefore, we investigated whether ASD and SAD traits are positively correlated with total dwell time at other facial areas.

Correlations between AQ and SAS scores and total dwell times on the AOIs are given in Table 9.2. As one participant did not complete the SAS questionnaire, correlations for the SAS scores are based on 64 instead of 65 participants. For both the AQ and SAS scores, significant negative correlations were observed for the total dwell time on the eyes AOI. When examining this per eye, total dwell time was negatively correlated with the total dwell time on the right eye AOI, but not the left eye AOI. Total dwell time on the nose AOI was positively correlated with both AQ and SAS scores. No other significant correlations were observed. This indicates that for participants scoring high on AQ or SAS looked less at the eyes and more at the nose of the other participant.

In order to get a clearer picture of what these correlations mean in terms of the total dwell times on the AOIs, groups were split by the median score on the respective questionnaire. Median AQ score was 14 (range 5-37). Participants with an AQ score equalling the median score or below were assigned to the low AQ group. Participants with an AQ score exceeding the median score were assigned to the high AQ group. The same was done for the SAS scores. Median SAS score was 44.5 (range 8-71). As can be seen in

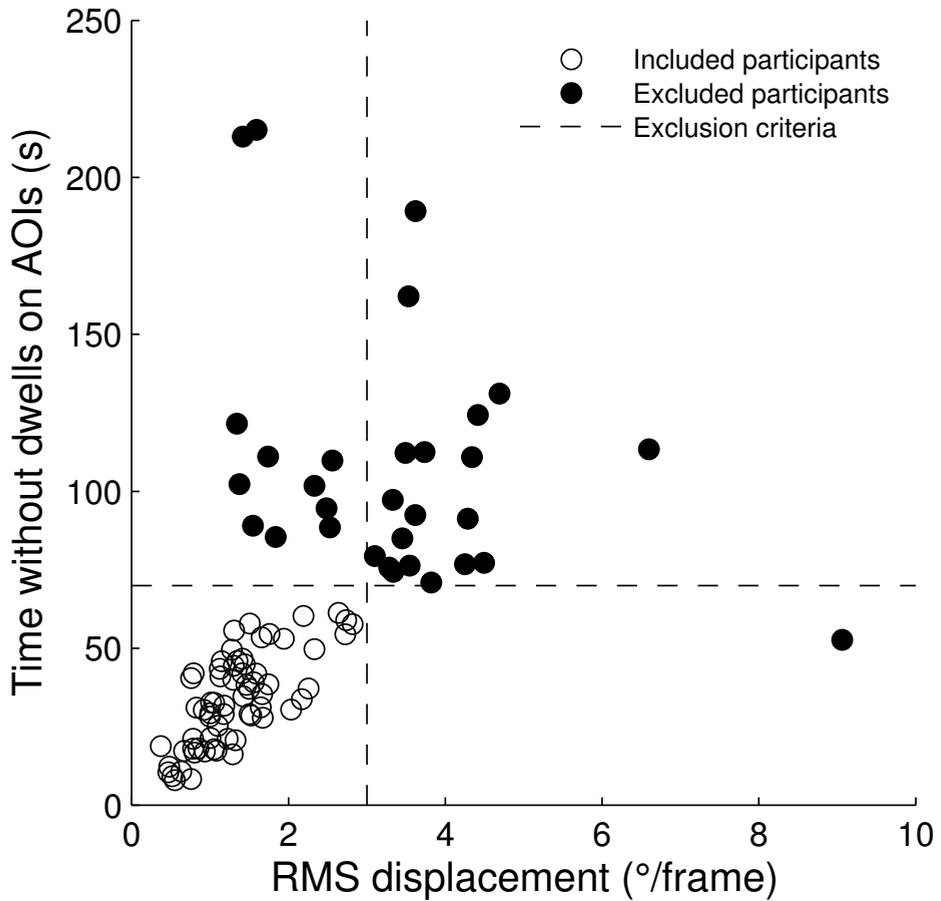


Figure 9.2.: Data quality measures for all participants. Each participant is depicted by a circle. Open circles represent included participants, closed circles excluded participants, based on the exclusion criteria marked with dashed lines.

## 9. Gaze in interaction in subclinical ASD and SAD

Table 9.2.: Correlations between AQ and SAS scores and total dwell times on AOIs.

AOI	AQ score	SAS score
Eyes	-0.25*	-0.26*
Left eye	-0.02	-0.02
Right eye	-0.31*	-0.32**
Nose	0.34***	0.33***
Mouth	0.16	0.05
Non	-0.17	-0.09
No eye-tracking data	-0.04	-0.07

\*  $p < 0.05$  \*\*  $p < 0.01$  \*\*\*  $p < 0.001$

Figure 9.3, participants scoring high on the AQ questionnaire looked less at the eyes and more at the nose compared to participants scoring low on the AQ questionnaire. This amounted to a difference of roughly 20 seconds on the nose and 30 seconds on the right eye, out of the total 300 seconds the experiment lasted. A similar pattern emerged for the participants scoring high on the SAS questionnaire, although the high SAS group looked up to 45 seconds less at the right eye compared to the low SAS group. This is nearly a sixth of the total time in the experiment, and between 40 and 50 % of the time the low SAS group spent looking at the right eye. As a shorter total dwell time can be the result of either less dwells, shorter dwells, or a combination of both, we looked at the correlations with number of dwells and mean dwell duration. This revealed that the shorter/longer total dwell times were the result of a combination of number and duration of individual dwells.

Given that the difference between the high and low AQ group looks very similar to that of the high and low SAS group, this was further investigated. First of all, the correlation between the AQ and SAS scores was 0.51 ( $p < 0.001$ ), which indicates that there was significant overlap between the two low and high groups, regardless of questionnaire. To get a clearer picture, the AQ score was split into its sub-scales (see section *Questionnaires*) and correlated with the total dwell time to the right eye and nose AOI as the

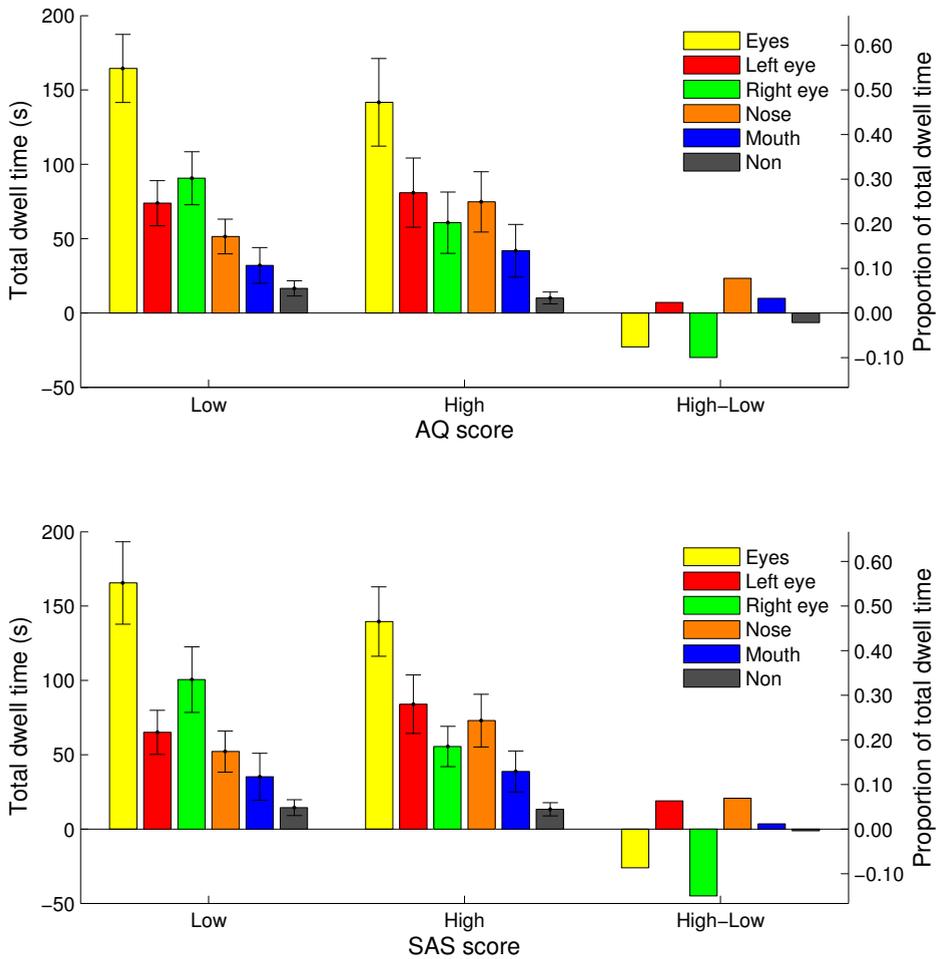


Figure 9.3.: Total dwell times to the AOI for the low and high AQ and SAS groups, split by the median questionnaire score. The mean difference in total dwell times to the AOIs between the high and low groups is depicted on the right. Error bars indicate 95% confidence interval of the mean.

## 9. Gaze in interaction in subclinical ASD and SAD

largest differences were observed here. As can be seen in Table 9.3, the differences in total dwell time to the right eye and nose AOI were significantly correlated only with the social skills and communication sub-scales. It may be that the social skills and communication sub-scales of the AQ measure some aspect that the SAS questionnaire also measures, and as such the overlap in the correlations with gaze behavior are not entirely surprising.

In sum, the one-way analysis revealed that gaze behavior to faces in interaction was significantly correlated with (sub)clinical psychopathology. As expected, AQ and SAS scores were negatively correlated with total dwell time the eyes, and particularly for the right eye. AQ and SAS scores were positively correlated with total dwell time to the nose AOI. Patterns were similar for the AQ and SAS scores, likely due to the fact that the relation with the AQ score was only in the social skills and communication sub-scales.

### 9.5.3. Paired gaze data

The paired gaze data were used to explore the relation between the paired gaze states (as defined in the *Methods* section) and ASD and SAD traits. Correlations between the total duration, mean duration and frequency of paired gaze states and the paired AQ and SAS scores are given in Table 9.4. There are several patterns that may be observed. First, paired AQ scores were negatively correlated with the total duration of two-way eye gaze, yet positively correlated with the total duration of one-way eye gaze. This pattern was further explored by splitting the groups by the median paired AQ score (pairs of observers scoring the median were assigned to the low paired AQ group). The median paired AQ score was 14.5 (range 6-27). In order to compare this pattern to SAS traits, the same was done for the paired SAS score. The median paired SAS score was 43 (range 27.5-68). As can be seen in Figure 9.4, the group with a high paired AQ score were engaged in two-way eye gaze for a shorter total duration (roughly 30 seconds), yet longer engaged in one-way eye gaze (roughly 30 seconds) compared to the low paired AQ group. This pattern was not observed for

Table 9.3.: Correlations between AQ sub-scale scores and total dwell times on right eye and nose.

AOI	Social skills	Attention switching	Attention to detail	Communication	Imagination
Right eye	-0.32**	-0.18	-0.10	-0.29*	-0.10
Nose	0.29*	0.15	0.21	0.30*	0.15

\*  $p < 0.05$ \*\*  $p < 0.01$

9. Gaze in interaction in subclinical ASD and SAD

Table 9.4.: Correlations between paired AQ and SAS scores and total duration, mean duration and frequency of paired gaze states.

Gaze state	AQ score	SAS score
Two-way total duration	-0.50*	-0.38
Two-way mean duration	-0.37	-0.53*
Two-way frequency	-0.34	-0.00
One-way total duration	0.51*	-0.21
One-way mean duration	0.30	-0.51*
One-way frequency	0.06	0.54*
No eye gaze total duration	0.17	0.60**
No eye gaze mean duration	0.18	0.21
No eye gaze frequency	0.20	0.69***

\*  $p < 0.05$  \*\*  $p < 0.01$  \*\*\*  $p < 0.001$

the high paired SAS group, which showed a shorter total time engaged in both one- and two-way eye gaze (roughly 10 seconds), and an increase in the total duration of no eye gaze.

While no other patterns were observed for the paired AQ score, there was for the paired SAS score. The paired SAS scores were negatively correlated with the mean duration of one-way eye gaze, yet positively correlated with the frequency of one-way eye gaze. As can be seen in Figure 9.5, the difference between the high and low paired SAS groups amounted to roughly 75 periods more of one-way eye gaze, yet of roughly 0.2 s shorter duration.

In sum, the two-way analysis revealed that paired gaze states were differently related to paired AQ and SAS scores. Pairs of observers scoring high on the AQ were engaged in less two-way eye gaze, yet more in one-way eye gaze compared to pairs of observers scoring low on the AQ. Pairs of observers scoring high on the SAS were more frequently engaged in one-way eye gaze, yet of shorter duration, compared to pairs of observers scoring low on the SAS.

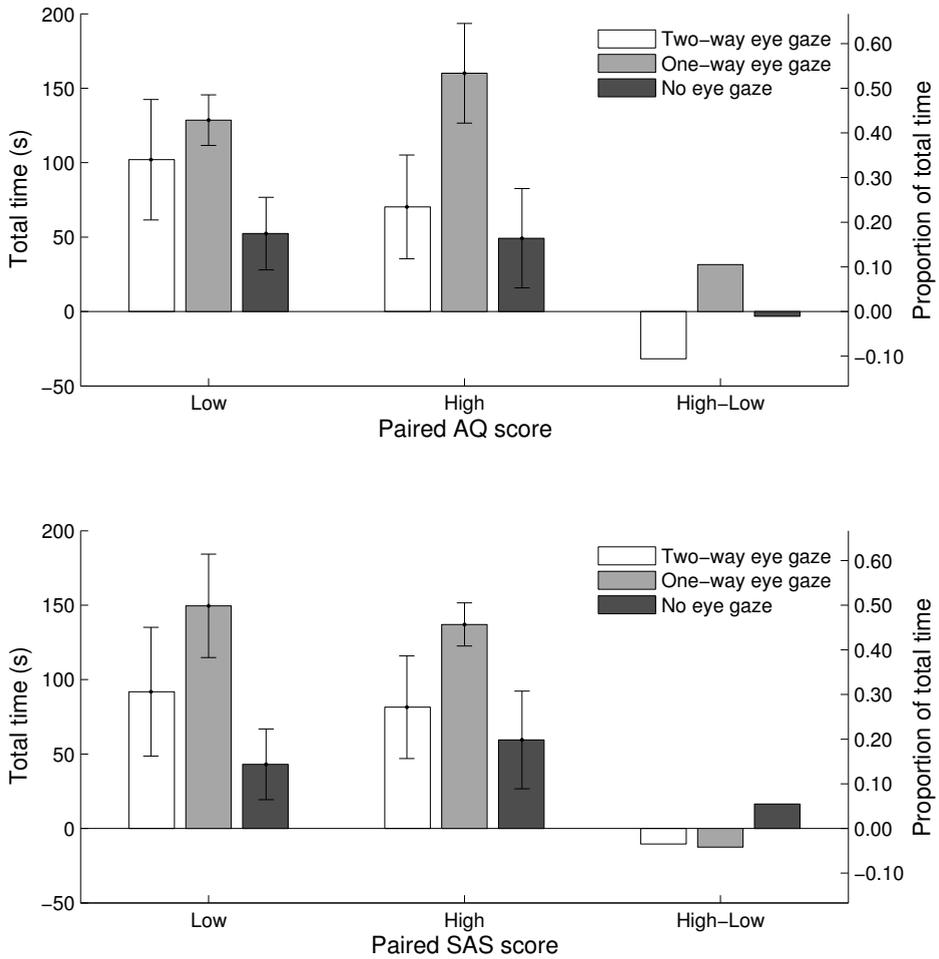


Figure 9.4.: Total duration for paired gaze states for low and high paired AQ and SAS groups, split by the median questionnaire score. The mean difference in total time for the paired gaze states between the high and low groups is depicted on the right. Error bars indicate 95% confidence interval of the mean.

## 9. Gaze in interaction in subclinical ASD and SAD

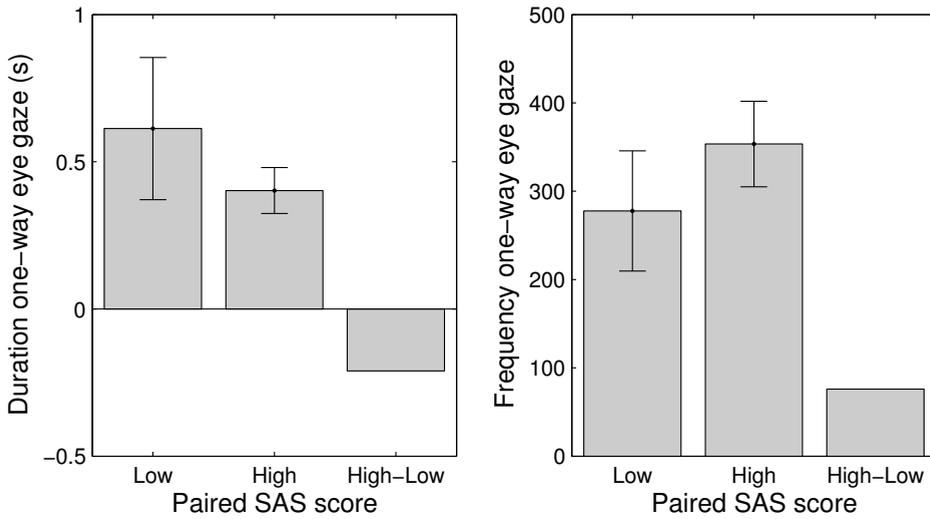


Figure 9.5.: Mean duration and frequency for one-way eye gaze for low and high paired SAS groups, split by the median SAS score. Error bars indicate 95% confidence interval of the mean.

## 9.6. Discussion

A novel social interaction dual eye-tracking setup, introduced by (Hessels et al., 2017), was used to investigate gaze behavior in dyadic interaction. Research over the last decades has tended to compartmentalize face processing in sub problems, and while this has resulted in a plethora of knowledge, questions have been posed about its generalizability to everyday social situations. This is particular relevant for gaze behavior in clinical psychopathology, where differences between clinical groups and controls have been inconsistent. By investigating the relation between Autism Spectrum Disorder (ASD) and Social Anxiety Disorder (SAD) traits in a general population and gaze behavior in interaction, we aim to assess the sensitivity of gaze behavior in social interaction to (sub)clinical psychopathology, and by extension the fruitfulness of this interactive approach to gaze behavior. The social interaction dual eye-tracking setup we employed allows for the registration of high-resolution gaze behavior to faces in interaction. Eye-tracking

data were analyzed in two ways. First, individual gaze data were used to answer the question where people look on the face of another in dyadic interaction, and how this is modulated by (sub)clinical psychopathology. We hypothesized that observers scoring high on ASD and SAD traits would show shorter total dwell times to the eye region, as evidenced by a negative correlation. Second, paired gaze data were used to explore how gaze behavior of a pair in interaction is related to (sub)clinical psychopathology. This was examined by exploring the relation between average ASD and SAD traits in a pair and the total time, frequency, and duration of paired gaze states (e.g. simultaneously looking at the eye region of the other).

The individual gaze analyses revealed that ASD and SAD traits were indeed negatively correlated with total dwell time at the eyes. While this was true for the area comprising both eyes, it appeared to be largely due to the total dwell time to the right eye. Moreover, both ASD and SAD traits were positively correlated with the total dwell time at the nose. Median split analyses revealed that the high AQ (higher scores indicate more ASD traits) group looked roughly 20 seconds longer on the nose and 30 seconds less on the right eye compared to the low AQ group, out of the total 300 seconds the experiment lasted. Moreover, the high SAS (higher scores indicate more SAD traits) group looked roughly 45 seconds less on the right eye compared to the low SAS group. As the pattern of differences in total dwell time between the low and high AQ group was similar to that of the low and high SAS group, we investigated this further. Correlation between total dwell times on the right eye and nose and the sub-scales of the AQ revealed that gaze behavior was related mainly to the social skills and communication sub-scales. Additionally, score on the AQ and SAS were positively correlated, indicating that both questionnaires might have tapped into similar traits. Concluding, the individual gaze data analysis corroborate the hypothesis that ASD and SAD traits modulate total dwell time to facial features in dyadic interaction, specifically for the eye and nose region.

## 9. Gaze in interaction in subclinical ASD and SAD

The main benefit of the eye-tracking setup employed was the ability to investigate the gaze behavior of pairs of observers. For the paired gaze analyses, three gaze states were defined: two-way eye gaze (both participants looking at the eyes of the other), one-way eye gaze (one participant looking at the eyes, but not the other) and no eye gaze (both participants not looking at the eyes of the other). Two main patterns emerged. First, high-AQ pairs of observers were engaged in less two-way eye gaze, yet more in one-way eye gaze compared to low-AQ pairs of observers. Second, high-SAS pairs of observers were more frequently engaged in one-way eye gaze, yet of shorter duration, compared to low-SAS pairs of observers. These findings are particularly relevant given several proposed models on gaze behavior in ASD and SAD.

Although there is still no consensus on whether fixation to the eyes is in fact reduced in ASD compared to controls (Senju & Johnson, 2009a), several models for this apparent lack of eye fixation have been proposed. Importantly, a distinction between ‘gaze aversion’ and ‘gaze indifference’ is being investigated, where the ‘gaze aversion’ model states individuals with ASD actively avoid the eye region of other and the ‘gaze indifference’ model states there is no particular avoidance, but mere omission of fixating the eye region (Moriuchi et al., 2017). Where the study by Moriuchi et al. (2017) corroborates the ‘gaze indifference’ model, the findings obtained here may be easier to explain in a ‘gaze aversion’ model. Pairs of observers scoring higher on the AQ were more often engaged in one-way eye gaze, yet less often in two-way eye gaze compared to low-AQ pairs of observers. If high AQ scores are only predictive of time spent looking at the eyes, but not of an avoidance of two-way eye gaze (which is sometime referred to as ‘eye contact’), one would expect high-AQ pairs of observers to show both diminished two-way and one-way eye gaze. It should be noted, however, that ASD is here investigated as a (sub)clinical trait in a general population, and not in a clinical group. Moreover, it is impossible to completely disentangle ASD from SAD traits here, given the positive correlation between the scores on the two questionnaires observed in the present study. Investigating gaze behavior in interaction may, however, provide valuable insights

in the context of these ‘gaze aversion’ and ‘gaze indifference’ models of eye fixation in ASD.

In SAD, a ‘vigilant-avoidance’ hypothesis is described that states that individuals with SAD are initially more attentive to ‘social threat cues’ compared to TD individuals, but hereafter actively avoid these ‘social threats’ (e.g. Horley et al., 2004). In the context of gaze behavior to faces, if ‘eye contact’ is considered a social threat, than individuals with SAD might be fast to fixate these cues (the ‘vigilance’), yet also fast to fixate elsewhere (the ‘avoidance’). In the present study, individuals scoring higher on SAD traits generally looked more often to the eyes while the other person did not, but did so for a shorter duration. Given that the ‘vigilant-avoidance’ hypothesis has not been specified as such for gaze behavior in interaction, the findings here are not interpreted as a corroboration of this hypothesis. However, it indicates that much more specific predictions may be investigated on the interplay of gaze between two partners, given a set of psychopathological traits.

## 9.7. Conclusion and limitations

A social interaction dual eye-tracking setup was used to investigate gaze behavior in interaction, and its relation to (sub)clinical psychopathology. We report that individuals scoring high on ASD and SAD traits look less at the eye region and more at the nose region of an interaction partner, corroborating long-standing findings obtained using static pictures and videos. Moreover, we are the first to report on the following findings on gaze behavior in interaction: first, pairs of observers scoring high on ASD traits were less engaged in two-way eye gaze, but more engaged in one-way eye gaze, compared to pairs of observers scoring low on ASD traits. Second, pairs of observers scoring high on SAD traits were more often engaged in one-way eye gaze, yet for shorter duration than pairs of observers scoring low on SAD traits. These findings provide intriguing possibilities for the investigation of gaze behavior in interaction in psychopathology. Moreover, the marked differences in gaze behavior between low and high-trait

## 9. *Gaze in interaction in subclinical ASD and SAD*

individuals attest to the sensitivity of gaze behavior in dyadic interaction to (sub)clinical psychopathology. Future research may benefit from more studies investigating gaze behavior in social settings.

Although there are promising findings reported here, several caveats are to be noted. First, a general population was assessed on (sub)clinical psychopathology. The question remains how this generalizes to gaze behavior in clinical groups. More specifically, future research may uncover how gaze behavior manifests itself within pairs observers from clinical groups, and between observers from clinical groups and typically developing individuals. Second, the individual gaze analyses were based on eye-tracking data that were also used in the paired gaze analyses. While the individual gaze analyses do allow for good comparison to research with static pictures or videos of faces, the findings from the individual gaze analyses cannot be 100% separated from the findings from the paired gaze analyses. Investigating gaze in interaction allows gaze behavior to be modeled as the output of a system composed of two people. However, disentangling these two individuals is a challenge. A first attempt has recently been made (Hessels et al., 2017), and we welcome research investigating the close interplay of gaze between interaction partners.

## **Acknowledgements**

This work was supported by a Netherlands Organization for Scientific Research (NWO) VICI grant (No. 45307004) awarded to CK and by the Consortium on Individual Development (CID). CID is funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the NWO (Grant No. 024.001.003).

## References

- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders*. American Psychiatric Association, Washington, DC, 5th edition.
- Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001a). The “reading the mind in the eyes” test revised version: A study with normal adults, and adults with asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry*, 42(2):241–251.
- Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001b). The autism-spectrum quotient (AQ): Evidence from asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders*, 31(1):5–17.
- Birmingham, E., Bischof, W. F., & Kingstone, A. (2008). Gaze selection in complex social scenes. *Visual Cognition*, 16(2-3):341–355.
- Broz, F., Lehmann, H., Nehaniv, C. L., & Dautenhahn, K. (2012). Mutual gaze, personality, and familiarity: Dual eye-tracking during conversation. *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*.
- Chawarska, K. & Shic, F. (2009). Looking but not seeing: Atypical visual scanning and recognition of faces in 2 and 4-year-old children with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, 39(12):1663–1672.
- Dalton, K. M., Nacewicz, B. M., Johnstone, T., Schaefer, H. S., Gernsbacher, M. A., Goldsmith, H. H., Alexander, A. L., & Davidson, R. J. (2005). Gaze fixation and the neural circuitry of face processing in autism. *Nature neuroscience*, 8(4):519–526.
- Damm, O., Malchus, K., Jaecks, P., Stenneken, P., Krach, S., Paulus, F., Becker, K., Jansen, A., Naber, M., Kamp-Becker, I., Einhaeuser-Treyer, W., & Wrede, B. (2013). Different gaze behavior in human-robot interaction in Asperger’s syndrome: an eye-tracking study. *2013 IEEE RO-MAN The 22nd IEEE International Symposium on Robot and Human Interactive Communication*.
- Dapretto, M., Davies, M. S., Pfeifer, J. H., Scott, A. A., Sigman, M., Bookheimer, S. Y., & Iacoboni, M. (2006). Understanding emotions in others: Mirror neuron dysfunction in children with autism spectrum disorders. *Nature neuroscience*, 9(1):28–30.
- Freeth, M., Foulsham, T., & Kingstone, A. (2013). What affects social attention? Social presence, eye contact and autistic traits . *PLOS One*, 8(1):e53286.

## 9. Gaze in interaction in subclinical ASD and SAD

- Frischen, A., Bayliss, A. P., & Tipper, S. P. (2007). Gaze cueing of attention: Visual attention, social cognition, and individual differences. *Psychological Bulletin*, 133(4):694–724.
- Gobel, M. S., Kim, H. S., & Richardson, D. C. (2015). The dual function of social gaze. *Cognition*, 136:359–364.
- Guillon, Q., Hadjikhani, N., Baduel, S., & Rogé, B. (2014). Visual social attention in autism spectrum disorder: Insights from eye tracking studies. *Neuroscience & Biobehavioral Reviews*, 42:279–297.
- Gullberg, M. & Holmqvist, K. (2006). What speakers do and what addressees look at: Visual attention to gestures in human interaction live and on video. *Pragmatics & Cognition*, 14(1):53–82.
- Henderson, J. M., Williams, C. C., & Falk, R. J. (2005). Eye movements are functional during face learning. *Memory & Cognition*, 33(1):98–106.
- Hernandez, N., Metzger, A., Magné, R., Bonnet-Brilhault, F., Roux, S., Barthelemy, C., & Martineau, J. (2009). Exploration of core features of a human face by healthy and autistic adults analyzed by visual scanning. *Neuropsychologia*, 47:1004–1012.
- Hessels, R. S., Kemner, C., van den Boomen, C., & Hooge, I. T. C. (2016). The area-of-interest problem in eyetracking research: A noise-robust solution for face and sparse stimuli. *Behavior Research Methods*, 48(4):1694–1712.
- Hessels, R. S., Cornelissen, T. H. W., Hooge, I. T. C., & Kemner, C. (2017). Gaze behavior to faces during dyadic interaction. *Canadian Journal of Experimental Psychology*.
- Hoekstra, R. A., Bartels, M., Cath, D. C., & Boomsma, D. I. (2008). Factor structure, reliability and criterion validity of the autism-spectrum quotient (AQ): A study in Dutch population and patient groups. *Journal of Autism and Developmental Disorders*, 38(8):1555–1566.
- Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press.
- Horley, K., Williams, L. M., Gonsalvez, C., & Gordon, E. (2003). Social phobics do not see eye to eye: A visual scanpath study of emotional expression processing. *Journal of anxiety disorders*, 17:33–44.
- Horley, K., Williams, L. M., Gonsalvez, C., & Gordon, E. (2004). Face to face: Visual scanpath evidence for abnormal processing of facial expressions in social phobia. *Psychiatry Research*, 127:43–53.
- Jones, W. & Klin, A. (2013). Attention to eyes is present but in decline in 2–6-month-old infants later diagnosed with autism. *Nature*, 504:427–431.

- Jones, W., Carr, K., & Klin, A. (2008). Absence of preferential looking to the eyes of approaching adults predicts level of social disability in 2-year-old toddlers with autism spectrum disorder. *Archives of General Psychiatry*, 65(8):946–954.
- Kingstone, A. (2009). Taking a real look at social attention. *Current Opinion in Neurobiology*, 19:52–56.
- Kingstone, A., Smilek, D., & Eastwood, J. D. (2008). Cognitive ethology: A new approach for studying human cognition. *British Journal of Psychology*, 99(3):317–340.
- Klin, A., Jones, W., Schultz, R., Volkmar, F., & Cohen, D. (2002). Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Archives of General Psychiatry*, 59:809–816.
- Laidlaw, K. E. W., Foulsham, T., Kuhn, G., & Kingstone, A. (2011). Potential social interactions are important to social attention. *Proceedings of the National Academy of Sciences*, 108(14):5548–5553.
- Langton, S. R. H. (2000). The mutual influence of gaze and head orientation in the analysis of social attention direction. *The Quarterly Journal of Experimental Psychology*, 53(3):825–845.
- Langton, S. R. H., Watt, R. J., & Bruce, V. (2000). Do the eyes have it? Cues to the direction of social attention. *Trends in Cognitive Sciences*, 4(2):50–59.
- Leppänen, J. M. & Nelson, C. A. (2009). Tuning the developing brain to social signals of emotions. *Nature Reviews Neuroscience*, 10:37–47.
- McPartland, J. C., Webb, S. J., Keehn, B., & Dawson, G. (2011). Patterns of visual attention to faces and objects in autism spectrum disorder. *Journal of Autism and Developmental Disorders*, 41(2):148–157.
- Merin, N., Young, G. S., Ozonoff, S., & Rogers, S. J. (2007). Visual fixation patterns during reciprocal social interaction distinguish a subgroup of 6-month-old infants at-risk for autism from comparison infants. *Journal of Autism and Developmental Disorders*, 37:108–121.
- Moriuchi, J. M., Klin, A., & Jones, W. (2017). Mechanisms of diminished attention to eyes in autism. *American Journal of Psychiatry*, 174(1):26–35.
- Moukheiber, A., Rautureau, G., Perez-Diaz, F., Soussignan, R., Dubal, S., Jouvent, R., & Pelissolo, A. (2010). Gaze avoidance in social phobia: Objective measure and correlates. *Behaviour Research and Therapy*, 48(2):147–151.
- Pelphrey, K. A., Sasson, N. J., Reznick, J. S., Paul, G., Goldman, B. D., & Piven, J. (2002). Visual scanning of faces in autism. *Journal of Autism and Developmental Disorders*, 32(4):249–261.

## 9. Gaze in interaction in subclinical ASD and SAD

- Rice, K., Moriuchi, J. M., Jones, W., & Klin, A. (2012). Parsing heterogeneity in autism spectrum disorders: Visual scanning of dynamic social scenes in school-aged children. *Journal of the American Academy of Child & Adolescent Psychiatry*, 51(3):238–248.
- Risko, E. F., Laidlaw, K. E. W., Freeth, M., Foulsham, T., & Kingstone, A. (2012). Social attention with real versus reel stimuli: Toward an empirical approach to concerns about ecological validity. *Frontiers in Human Neuroscience*, 6(1):143.
- Rutherford, M. D. & Towns, A. M. (2008). Scan path differences and similarities during emotion perception in those with and without autism spectrum disorders. *Journal of Autism and Developmental Disorders*, 38(7):1371–1381.
- Senju, A. & Johnson, M. H. (2009a). Atypical eye contact in autism: models, mechanisms and development. *Neuroscience & Biobehavioral Reviews*, 33:1204–1214.
- Senju, A. & Johnson, M. H. (2009b). The eye contact effect: mechanisms and development. *Trends in Cognitive Sciences*, 13(3):127–134.
- Smilek, D., Birmingham, E., Cameron, D., Bischof, W., & Kingstone, A. (2006). Cognitive ethology and exploring attention in real-world scenes. *Brain Research*, 1080:101–119.
- Speer, L. L., Cook, A. E., McMahon, W. M., & Clark, E. (2007). Face processing in children with autism: Effects of stimulus contents and type. *Autism*, 11(3):265–277.
- Sterling, L., Dawson, G., Webb, S., Murias, M., Munson, J., Panagiotides, H., & Aylward, E. (2008). The role of face familiarity in eye tracking of faces by individuals with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, 38(9):1666–1675.
- Van der Geest, J. N., Kemner, C., Camfferman, G., Verbaten, M. N., & van Engeland, H. (2002a). Looking at images with human figures: Comparison between autistic and normal children. *Journal of Autism and Developmental Disorders*, 32(2):69–75.
- Van der Geest, J. N., Kemner, C., Verbaten, M. N., & van Engeland, H. (2002b). Gaze behavior of children with pervasive developmental disorder toward human faces: A fixation time study. *Journal of Child Psychology and Psychiatry*, 43(5):669–678.
- Võ, M. L. H., Smith, T. J., Mital, P. K., & Henderson, J. M. (2012). Do the eyes really have it? Dynamic allocation of attention when viewing moving faces. *Journal of Vision*, 12(13):3.
- Wieser, M. J., Pauli, P., Alpers, G. W., & Mühlberger, A. (2009). Is eye to eye

## 9.7. Conclusion and limitations

contact really threatening and avoided in social anxiety?—An eye-tracking and psychophysiology study. *Journal of anxiety disorders*, 23:93–103.

Willems, L. F. M., Tuender-De Haan, H. A., & Defares, P. B. (1973). Een schaal om sociale angst te meten. *Nederlands Tijdschrift voor de Psychologie*, 28: 415–422.

Yarbus, A. L. (1967). *Eye movements and vision*. Plenum Press, New York.



## 10. Discussion

## 10. Discussion

The goal of the present dissertation was to explore two possible early markers of Autism Spectrum Disorder (ASD): visual search superiority, and gaze behavior during face perception. These possible markers were explored as they pertain to both the social deficits in ASD (gaze during face perception) and visual peculiarities (visual search superiority), thereby tackling both diagnostic criteria for ASD. Moreover, visual search superiority and gaze during face perception have been widely studied with older children and adults with ASD. The main reasons for investigating early markers of ASD was to get a better understanding of the atypicalities in ASD as they emerge during development (e.g. Karmiloff-Smith, 2012), and to enable earlier diagnosis and intervention, as the average age of diagnosis is still around 6 to 7 years of age (van Berckelaer-Onnes et al., 2015).

Although exploring possible early markers of ASD was the main focus at the outset of the research conducted in this dissertation, it quickly became clear that the available methodology was critically lacking. As both early markers for ASD described above are dependent on eye-tracking technology, we first examined the available methodology. For example, as infants cannot be instructed and are difficult to restrict in movement, we expected that adult eye-tracking methodology with regard to experimental setup and data processing might not suffice for infant participants. Second, investigating early markers for ASD not only depends on adequate methodology, but also on knowledge of typical development. With respect to one of the possible early markers, visual search superiority, relatively little knowledge on typical development of visual search behavior was available. Third, the conclusions drawn from studies investigating the other possible early marker for ASD, gaze behavior during face perception, have been highly inconsistent: Some studies concluded that individuals with ASD look less at the eye region, whereas other studies did not. Therefore, a novel approach for investigating gaze behavior during face perception was developed by investigating gaze behavior in dyadic interaction. This also meant that new methodology had to be developed to investigate gaze behavior in interaction. As such, a large part of the dissertation was focused on developments making future investigations of early markers in ASD possible.

This chapter is structured in three parts. In the first section (*Advances in eye-tracking methodology*), the state of eye-tracking methodology in developmental psychology is examined, and the advances in eye-tracking methodology made in the present dissertation are discussed. In the second section, *Visual search in ASD* is addressed. The body of literature on visual search behavior in ASD can briefly be summarized as follows: Visual search superiority in ASD has been observed across a wide age range, from toddlerhood to adulthood, and is generally more apparent when difficult search targets are used. We investigated visual search superiority in ASD, and discuss the limit and scope of visual search superiority in ASD. Moreover, we investigated visual search behavior in infancy as a benchmark for comparing infants at-risk for ASD. The main findings from these studies are summarized and discussed. In the third section, *Gaze behavior during face perception in ASD* is addressed. We investigated adult gaze behavior in dyadic interaction, and its relation to (sub)clinical psychopathological traits, including ASD traits. The main findings from these studies are summarized and discussed.

### 10.1. Advances in eye-tracking methodology

For a typical eye-tracking study, there is a long chain of steps from experimental setup to data analysis that one goes through.

1. A suitable eye tracker is chosen, and a participant is carefully positioned in front of it.
2. The eye tracker signal is calibrated and validated by having a participant look at a number of points on screen.
3. Visual stimuli may be presented during which eye-tracking data is collected.
4. In order to draw conclusions about where one has looked on a visual stimulus, parts of the eye-tracking signal are classified as saccades and fixations. These fixations are subsequently mapped to the visual stimulus.

## 10. Discussion

5. After fixations are mapped to the visual stimulus, measures such as the number of area of interest (AOI) hits or total time spent looking at an AOI are calculated. Statistical analyses may subsequently be carried out on these outcome measures.

While all these steps have been investigated quite thoroughly for adult participants (Holmqvist et al., 2011), relatively little attention has been given to the problems young children and infant participants may introduce for this chain. While the advantages of using eye-tracking technology in infant research are evident (Aslin & McMurray, 2004; Oakes, 2012), the disadvantages are only recently being hinted at (Wass et al., 2013, 2014). For example, eye-tracking data from infant research are typically of lower quality than that of adult research. Remedying such problems requires signal processing skills that previous techniques for studying infant gaze behavior (such as observational techniques) did not require. Moreover, only recently has a leading developmental journal (*Infancy*) required that eye-tracking methodology be clearly stated in its submissions (Oakes, 2010). Prior to this, no requirements were made at all on what to report on eye-tracking methodology. As a large number of choices are made in the chain from experimental setup to data analysis, knowledge is desperately needed on exactly how these choices affect eye-tracking data quality and the outcome measures. In the present dissertation, all steps in the chain above are investigated.

It has previously been reported that eye-tracking with infants may lead to lower data quality compared to adult participants (Wass et al., 2014), which poses a problem for data analysis (Wass et al., 2013). Given the inadequacy of the automatic analysis tools available for eye-tracking data of the quality observed in infant eye-tracking research, researchers have even turned to manual correction of eye-tracking data analysis (Saez de Urabain et al., 2015). This is particularly unwanted, given that eye-tracking methodology has increasingly been used over observational techniques for its objectivity. It seems that the chain of steps from experimental setup to data analysis in infant eye-tracking research needs better argumentation. What choices

are made, and why? How do these choices affect data quality and outcome measure of eye-tracking studies?

In **Chapter 2** we discussed how one may make a choice of eye tracker for research with difficult participant groups (step 1 in the chain). We investigated 8 different eye-tracking setups with regard to their tracking recovery after a participants looks away from, and back to, the screen, and eye-tracking performance during non-optimal head orientations. Non-optimal head orientations are all orientations that differ from sitting upright in front of the eye tracker. Participants were adults who mimicked behaviors typically seen in infant eye-tracking research. We report that eye trackers vary in their robustness to non-optimal head orientations: loss of eye-tracking data differs widely between eye trackers. Moreover, systematic offsets – the difference between the actual gaze position of a participant and the reported gaze position by the eye tracker – differ widely too when eye trackers still report data. We suggest that eye-tracking researchers should test their eye trackers in the situations in which they are going to be used, rather than only depending on manufacturer specifications on data quality obtained in optimal conditions.

After a choice of eye tracker has been made, the next step would be to obtain high quality eye-tracking data (step 2 and 3 in the chain). High quality eye-tracking data contain small to no systematic error (or high accuracy), small to no variable error (high precision, or low noise level) and little to no data loss. In **Chapter 3** we investigated how eye-tracking data quality in infant research is affected by eye color, positioning (whether an infant is in a high chair or on the parents lap), measurement characteristics, and infant contentedness. We report that bluish eye color resulted in data of lower quality across all measures of data quality compared to brownish eye color. Moreover, data loss level was progressively higher as time in the experiment increased. Finally, we present tentative arguments that positioning affects accuracy of the measurement. We highlight that it is important that factors such as eye color are taken into account when comparing groups, such that differences in data quality are not confounding

## 10. Discussion

between-group comparisons.

In **Chapter 4**, we introduced a new fixation-detection algorithm, built specifically to achieve automatic analysis of eye-tracking data of varying noise and data loss levels (step 4 in the chain). The Identification by two-means clustering (I2MC) algorithm was compared against 7 state-of-the-art event-detection algorithms (algorithms that detect events such as e.g. fixations and saccades). We report that the output of the I2MC algorithm is most robust to high noise and data loss levels compared to the other algorithms. This is particularly relevant given that previous research has indicated that the choice of parameters in fixation-detection algorithms may actually reverse differences between ASD and TD children (Shic et al., 2008). Using the I2MC algorithm, such differences on the basis of data quality are no longer expected, provided that the data quality is within a reasonable noise and data loss range. Moreover, hand correction of fixation detection is no longer necessary (Saez de Urabain et al., 2015), as the I2MC algorithm works fully automatic.

In **Chapter 5**, we investigated different AOI production methods for face stimuli (step 4 and 5 in the chain). We report that AOI-based measures did not differ dramatically between AOI production methods. However, we report that AOIs implemented using the Voronoi or Limited-Radius Voronoi Methods (LRVT) are the most objective of researcher-defined AOIs with regard to the visual stimulus. This means that the least amount of subjective choices in terms of position, size and shape have to be made for AOIs using these methods. Moreover, the Voronoi and LRVT methods allow automatic implementation using a computer script. As they can be implemented automatically, they can significantly reduce investigator time when applied to video stimuli compared to when manual AOI construction is applied (e.g. AOIs are manually constructed per video frame). Finally, the AOI-based measures were most robust to increased noise when the Voronoi method or LRVT method with large radii were used. We conclude that large AOIs might be most noise-robust in sparse stimuli in general. When working with participant groups that might produce low quality

eye-tracking data, such as infants, large AOIs are therefore preferable.

The studies on eye-tracking methodology outlined in this dissertation have provided much better argumentation for all steps in the chain described above, from the choice of eye tracker, to setting up an eye-tracking measurement, and analyzing low-quality eye-tracking data from infant participants. However, this does not mean that eye-tracking methodology for infant research has now crystallized into a standard. There are several reasons why this is not the case. First, manufacturers are constantly working on new eye trackers, which improve both in terms of data quality in optimal, as well as in non-optimal conditions, over previous models. It is vital to assess new eye trackers with regard to their performance in non-optimal conditions in order to keep track of which eye trackers are suited for difficult participant groups, and which are not. We have already undertaken such new assessments (Niehorster et al., 2017). Second, the present chapters on eye-tracking data analysis have been restricted to the detection of fixations, and construction of AOIs for face and sparse stimuli. Clearly, infant eye-tracking research is not limited to these situations and variables. Future research should endeavor to develop eye-tracking analysis tools suited for infant eye-tracking data in a diverse range of stimuli and when different eye movements are of interest. Finally, investigating the impact of eye-tracking data quality across development is essential for longitudinal studies where eye-tracking data is being used to characterize individuals, for example in terms of (a)typical eye-movement behavior.

## 10.2. Visual search in ASD

As described in the **Introduction**, a large number of studies have reported visual search superiority for individuals with ASD compared to controls. In **Chapter 6** a visual search experiment was conducted with adults with ASD. However, contrary to our expectation, we observed no visual search superiority in the ASD group compared to the typically-developing (TD) group. As this experiment was based on an earlier study from our laboratory in which visual search superiority for the ASD group was reported

## 10. Discussion

(Kemner et al., 2008), the specific methodological differences between that study and the present experiment were considered. In the previous study by Kemner et al. (2008), the visual search displays contained a larger difference in angle between the target (a vertical line) and the non-targets (tilted lines). When the difference between target and non-targets is larger, the target can be located faster. Moreover, multiple set sizes were used in that study as opposed to one set size in the present experiment. Therefore, a second experiment was conducted which matched Kemner et al. (2008) in terms of the precise methodology. This time, visual search superiority was observed for the ASD group. Based on these observations, we conjectured that the perceptual load – the concept of how much load is being placed on the perceptual system – of a visual search display is critical to observing differences in visual search performance between individuals with ASD and TD individual. Specifically, we hypothesized that “... *as long as perceptual capacity leaves room for processing of additional task-(ir)relevant information, individuals with ASD will show superior performance. When perceptual load is so high that no room is left for processing additional task-(ir)relevant information, even in individuals with ASD, no differences will be found on static stimuli between individuals with ASD and controls.*”

The conjecture that perceptual load is critical in observing visual search superiority in ASD may be tested in a number of ways. First, when visual search displays are systematically varied from low to high perceptual load, one would expect to observe visual search superiority in ASD only for the medium load displays. On low and high perceptual load display, no observable differences in visual search performance between individuals with ASD and controls are to be expected. This would either be because target detection would occur almost instantaneous for both individuals with ASD and controls (in low perceptual load displays) or because target discrimination would have to be made on an item-per-item basis (in high perceptual load displays). For the latter, similar performance for individuals with ASD and controls would be expected, given that no superiority in target discrimination for individuals with ASD has been observed. Second, one might suggest that if visual search superiority depends on perceptual ca-

capacity for processing (ir)relevant information, the functional field of view (FFOV) – the part of the visual field from which information can be extracted – in a visual search task would be larger for individuals with ASD compared controls. However, Song et al. (2015) report that the FFOV of individuals with ASD is actually narrower compared to controls in a number or letter detection and identification task. Moreover, in an experiment with briefly presented search displays, performance as a function of target eccentricity did not show a different pattern for individuals with ASD compared to controls (Shirama et al., 2016). Critically, however, the FFOV of individuals with ASD and controls has never been investigated in a saccadic search task – when the target cannot be located at first glance, and an active exploration of the visual scene is necessary to locate the target.

### 10.2.1. Visual search in ASD: a perspective

While we framed the empirical findings on visual search superiority in ASD in **Chapter 6** in terms of perceptual load theory, a perspective for future research warrants an examination of other models on visual search superiority in ASD. Can the cause or mechanism of visual search superiority in ASD be accurately pinpointed? And what knowledge is critically lacking?

Kaldy et al. (2016) highlight two prevailing strands of models on visual search superiority in ASD. The first strand of models considers ASD superiority in search to be perceptual, either due to ‘enhanced perceptual discrimination’ or ‘enhanced perceptual functioning’ in individuals with ASD. However, Kaldy et al. (2016) state that these models are generally put forward in studies by process of elimination of other models, and that no study has reported a relation between discrimination ability and visual search performance. Enhanced perceptual discrimination thus seems unlikely an explanation of visual search superiority in ASD, and Kaldy et al. (2016) favor the second strand of models over perceptual explanations of visual search superiority in ASD.

## 10. Discussion

The second strand of models consider visual search superiority in ASD to be attentional. Kaldy et al. (2016), for example, propose that the individuals with ASD have ‘over-focused attention’, which increases performance on tasks that benefit from focused attention, including visual search tasks. The model is unclear, however, concerning what specific predictions may be made about visual search performance and behavior. Perhaps this model suggests that information is more efficiently extracted by individuals with ASD compared to controls? In that case, one might expect shorter fixation durations in individuals with ASD, or fewer fixations on targets that have already been fixated. The first suggestion has been investigated by two studies on eye movements in visual search. These have either reported that fixation durations are shorter for individuals with ASD compared to controls (Joseph et al., 2009), or that fixation durations do not differ between individuals with ASD and controls (Kemner et al., 2008). As no clear testable hypotheses were derived from the model, however, it cannot be corroborated or falsified at this point.

Keehn et al. (2013) disqualify the attentional model proposed by Kaldy et al. (2016), and propose an alternative attentional model of visual search superiority in ASD. They propose that the critical attentional impairment in ASD is in orienting behavior, which comprises disengaging, shifting, and reengaging attention. However, it would seem surprising to observe visual search superiority, given that visual search is comprised of consecutive shifts of attention across the visual scene when a target cannot be located at first glance (i.e. saccadic search). Concluding, while it might seem that perceptual models are discounted, attentional models by no means have filled the void and explained visual search superiority in ASD nor provided evidence in favor of them by newly tested hypotheses. Whether the perceptual load model poses a suitable alternative is for future research to determine. However, given the inadequacy of both perceptual and attentional models in explaining visual search superiority and providing testable hypotheses in terms of eye-movement behavior, a step back seems appropriate. If we understand the specific behavioral differences during visual search between individuals with ASD and controls, we should be able to generate better

models.

One way to better understand the fine-grained behavioral differences in visual search between individuals with ASD and controls is to use eye tracking to investigate eye movements during visual search. Only a few studies on visual search behavior in ASD have been conducted using eye tracking so far (Kemner et al., 2008; Joseph et al., 2009; Keehn & Joseph, 2016). Keehn & Joseph (2016) recently conducted a thorough investigation of visual search behavior in ASD using eye tracking. They investigated whether visual search superiority in ASD was due to better selection of possible target-candidates from the periphery or due to better discrimination of the target. However, they reported that individuals with ASD had shorter reaction times on trials without a target present compared to controls, and no differences in target discrimination or peripheral selection were observed. A number of previous studies (see **Introduction**) have, on the other hand, reported search superiority for target present trials also. It may therefore be that differences between individuals with ASD and controls in peripheral selection or target discrimination will be observed when the difference in search performance between groups is larger. The key finding Keehn & Joseph (2016) reported, was that individuals with ASD did not have a bias to go leftward first in the display, whereas controls did. They conjecture that the absence of such a bias may lead to faster processing of the scene if it is combined with a greater perceptual capacity for processing (ir)relevant information (cf. perceptual load theory). Although Keehn & Joseph (2016) have made a first step towards understanding visual search behavior in ASD, it remains unclear whether specific differences in visual search behavior between individuals with ASD and controls may be identified. Investigating eye movements during visual search – or saccadic search – across development, from infants to adults, may shed light on this question. It may thereby be a particularly useful approach in the search for an early marker for ASD.

### 10.2.2. Visual search as an early marker for ASD

As a stepping stone towards visual search as an early marker for ASD, we investigated saccadic search behavior in typically developing infants in **Chapter 7**. Before saccadic search in infants at-risk for ASD can be investigated, we need to know whether TD infants demonstrate visual search behavior in the absence of instructions. Moreover, investigating atypical saccadic search behavior requires knowledge of the characteristics of typical search behavior. Finally, if saccadic search behavior is to be used as a possible early marker for ASD, measures derived from this behavior should be reliable. The main findings from the investigation of saccadic search behavior in TD infants, how it can be characterized, and the test-retest reliability of saccadic search performance are discussed below.

We report that 10-month old infants search for a discrepant item in a visual display in the absence of instruction. This is concluded based on the fact that saccadic search performance was dependent on target and non-target dissimilarity, and that search performance was above chance. Next, infant saccadic behavior in search displays was characterized and compared to adult saccadic search behavior. Important similarities between infant and adult saccadic search behavior were observed, which may be summarized as follows. The initial fixation in a search display, which may be considered as the latency to initiate the first saccade, was longer than following fixations. Moreover, saccades generally followed the current trajectory with the same saccadic amplitude, or saccades returned to the previously fixated location. In other words, infants' saccades tended to either go in the same direction, or back to the location that was just fixated. Finally, test-retest reliability of oculomotor characteristics (e.g. fixation duration, and latency to initiate an eye movement) and saccadic search performance was examined. Oculomotor characteristics showed moderate to high test-retest reliability. Saccadic search performance – defined as average time infants took to fixate the search target – was moderately reliable. However, this was only the case for one of the conditions: where the search target was a 60°tilted line between vertical non-targets. Saccadic

### 10.3. Gaze behavior during face perception in ASD

search performance for the other conditions (a 30° or 90° titled line among vertical non-targets) was not reliable at all. The fact that test-retest reliability for saccadic search performance is lower than that of oculomotor characteristics may be due to the nature of the data. Depending on the direction of the first saccade, a target may be located quickly or not – i.e. if the first saccade happens to be in the direction of the saccade the target is located faster than when a first saccade is made in the opposite direction. This means that a reliable average time to target hit (as a measure for search performance) may require a large number of trials. Why search performance was only reliable for the 60° search target and not the other two search targets remains unknown.

Although we have shown that 10-month-old infants demonstrate spontaneous saccadic search behavior which is similar to adult search behavior in a number of ways, several questions remain. First, how does saccadic search behavior develop from birth to toddlerhood? Do infants search for discrepant items in the absence of instructions prior to 10 months? Investigating such questions is important if we want to draw conclusions on typical development of saccadic search behavior. Moreover, they are vital if atypical saccadic search behavior in infancy is to be distinguished from typical development, as has been observed in older children with ASD.

### 10.3. Gaze behavior during face perception in ASD

As stated in the **Introduction**, recent reviews on gaze behavior during face perception in ASD have concluded 1) that the assumption that “*individuals with ASD exhibit excess mouth and diminished eye gaze compared to TD individuals*” receives little support (Guillon et al., 2014, p. 286) and 2) that “*the available evidence at present suggests that the reduced fixation on the eyes in ASD is most prominent under conditions of high cognitive demand*” (Senju & Johnson, 2009, p. 1211). This may be surprising if one considers the marked impairments in social communication and interaction prevalent in ASD – of which abnormalities in eye contact is one aspect. However, upon reflecting on how most of these laboratory studies

## 10. Discussion

were conducted, this summary of the research on gaze behavior during face perception in ASD may make more sense. In most studies, non-responsive stimuli (e.g. pictures or videos of others) have been used to study how individuals with ASD look at faces. The question remains whether gaze behavior to a non-responsive stimulus may generalize to gaze behavior in social interaction. If the goal is to generalize gaze behavior toward non-responsive stimuli to social interaction, Brunswik's model of "representative design" states that the conditions of an experiment must represent those to which one wants to generalize (see Hammond, 1998, for an extensive discussion). The crucial condition here is whether a stimulus is responsive to the behavior of a participant. The concern for this lack of generalizability from non-responsive stimuli to social interaction is emphasized by a recent approach, "Cognitive ethology" (Smilek et al., 2006; Kingstone et al., 2008; Kingstone, 2009), which advocates studying behavior in natural settings before moving to the laboratory instead of vice versa.

In order to investigate gaze in interaction with the control of a laboratory study, we developed a dual eye-tracking interaction setup capable of recording eye movements with high temporal and spatial resolution. In the present dissertation, two studies were carried out using this novel dual eye-tracking interaction setup, which allowed the co-registration of eye movements from two participants while they could engage in dyadic interaction. In **Chapter 8**, we investigated whether gaze behavior to faces during dyadic interaction was similar to what had previously been observed in a plethora of studies using static pictures or videos. Specifically, we questioned whether there was still a preference for fixating the eyes when the person being looked at was an actual person, not a non-responsive representation. We corroborated the findings that people have a preference for fixating the eye region of others. Moreover, there was tentative evidence that gaze behavior to faces was coupled between participants, such that the total time spent looking at the eyes was highly correlated between the two partners in a dyad. However, when we explicitly manipulated the gaze behavior of a confederate in each dyad by giving instructions on where to look through an earpiece, no effects on the gaze behavior of the other were observed. If a coupling exists

between the gaze behavior of two persons in a dyad, it appears not easily to manipulate.

In **Chapter 9**, we investigated whether gaze behavior to faces during dyadic interaction was related to (sub)clinical psychopathological traits: specifically ASD and Social Anxiety Disorder (SAD). Based on previous research using static pictures or videos in these clinical groups, we expected that participants scoring high on ASD and SAD traits would look less at the eye region compared to participants scoring low on these traits. Note that although studies using static pictures and videos in individuals with ASD have drawn competing conclusions on gaze abnormalities to faces in ASD, we did expect gaze behavior to be related to ASD traits based on Senju & Johnson (2009). We report that ASD and SAD traits were indeed related to the total time spent looking at the eyes: participants scoring high on these traits looked less at the eyes of the other compared to participants scoring low on these traits. As ASD and SAD traits were positively correlated (participants scoring high on ASD traits also scored high on SAD traits), and very few people scored low on ASD and high on SAD traits (or vice versa), it was impossible to separate the relation between ASD and SAD traits and gaze behavior completely. However, when pairs of participants were considered on the basis of their average ASD or SAD traits, patterns for ASD and SAD diverged markedly. First, pairs of participants scoring high on ASD traits were engaged in one-way eye gaze (one participant looks at the eyes, whereas the other does not) for a longer total duration, yet were engaged in two-way eye gaze (both participants look at the eyes of the other) for a shorter total duration than pairs of participants scoring low on ASD traits. Pairs of participants scoring high on SAD traits, on the other hand, were more often engaged in one-way eye gaze, yet of shorter average duration, than pairs of participants scoring low on SAD traits.

There are a number of important points that we have learned from these two studies. First, we have reported a number of critical corroborations of research on gaze behavior to faces and its relation to (sub)clinical psychopathology. Moreover, these corroborations were obtained in an entirely

## 10. Discussion

novel setup where dyadic interaction occurs, as opposed to a static environment with non-responsive stimuli. Second, we have reported that pair-based measures of gaze are differentially related to ASD and SAD traits. One aspect in which further developments are necessary is analyzing the coupling between gaze of two persons more closely in time (as discussed in **Chapter 8**). For example, what happens when one person switches from looking at the eyes to the mouth of the other? Does the other follow? Or does the transition from eyes to mouth trigger a different response? Does it trigger a response at all? And if we find that such a coupling exists, how may it be affected in ASD? With the advance of our setup, other questions may be posed as well. How is one's gaze behavior to another person affected by that person being pre-recorded or not? How is gaze in interaction affected when a delay is introduced between the two interacting partners? The novel setup introduced here opens up a plethora of research avenues in which such questions may be answered.

One critical question is whether the dual eye-tracking interaction setup introduced in this dissertation is the way to gaze behavior during face perception as an early marker for ASD. A recent study suggests that it might be. Edmunds et al. (2017) were interested in measuring eye contact between children with ASD and social partners. They reported that both scoring from video observations, and mobile eye tracking (i.e. eye-tracking glasses), have not allowed researchers to go beyond a “looks at face” - “does not look at face” distinction, as the spatial resolution is too low. Their solution was to place a point-of-view camera on the social partner, and code from this video when the child with ASD looked them in the eyes. Only using this method did they observe differences between children with ASD and controls based on their gaze behavior, not with observations from stationary video cameras. With our interaction setup, however, it becomes possible to investigate gaze between children with or without ASD and their social partners with such high resolution that “eye contact” can be measured objectively. Moreover, much more fine-grained analyses of gaze in interaction become possible, which opens up new perspectives for research on (a)typical gaze during face perception.

## 10.4. Conclusions

Two potential early markers for ASD were explored: visual search behavior, and gaze during face perception. When the studies described in this dissertation were started, several requirements for the investigation of these markers in early development were lacking. For that reason, eye-tracking methodology for infant research was first critically examined, as both markers depend on eye-tracking technology. Second, visual search behavior of adults with ASD was not fully understood and knowledge on typical saccadic search behavior in infancy was lacking. Third, inconsistent conclusions drawn from studies on gaze behavior to faces in ASD led us to develop a novel approach for investigating gaze behavior during face perception.

The methodological advances introduced in **Chapters 2-5** have resulted in a better argumentation for the choice of eye tracker, achieving high data quality, and eye-tracking data analysis in infant research. These findings were necessary 1) in order to make the investigation of saccadic search in infancy using eye-tracking methodology possible and 2) in order to make automatic data analysis of gaze behavior in interaction possible. Visual search superiority in ASD was examined in **Chapter 6** and discussed in context of prevailing models of visual search superiority in ASD. Both perceptual and attentional models fall short of explaining search superiority in ASD or providing testable hypotheses, and an investigation of the fine-grained behavioral differences in saccadic search between individuals with ASD and controls across all ages was proposed. Saccadic search behavior in typically-developing infants was investigated in **Chapter 7**. Infants searched for discrepant items in the absence of instructions, and saccadic search behavior was similar to that observed in adults. A novel method for investigating gaze behavior during face perception was introduced in **Chapter 8**, and gaze behavior in dyadic interaction was linked to ASD and SAD traits in **Chapter 9**. Gaze behavior to faces in dyadic interaction revealed important corroborations of research using non-responsive stimuli. Moreover, gaze behavior during dyadic interaction revealed diverg-

## 10. Discussion

ing relations to ASD and SAD traits.

In conclusion, the present dissertation demonstrates important methodological advances that are crucial in order to grasp typical and atypical gaze behavior both early in development and in adults. These advances are paramount for our understanding of the development of saccadic search behavior from infancy to adulthood, and in order to grasp the atypicalities in ASD. Moreover, a novel setup was introduced for the investigation of gaze behavior during dyadic interaction, which has already revealed intriguing links between gaze behavior and (sub)clinical psychopathology. Our setup opens up entirely new avenues for the investigation of (a)typical gaze behavior in social interaction, and may help us grasp the core social deficits in ASD.

## References

- Aslin, R. N. & McMurray, B. (2004). Automated corneal-reflection eye tracking in infancy: Methodological developments and applications to cognition. *Infancy*, 6(2):155–163.
- Edmunds, S. R., Rozga, A., Li, Y., Karp, E. A., Ibanez, L. V., Rehg, J. M., & Stone, W. L. (2017). Brief report: Using a point-of-view camera to measure eye gaze in young children with autism spectrum disorder during naturalistic social interactions: A pilot study. *Journal of Autism and Developmental Disorders*, pages 1–7.
- Guillon, Q., Hadjikhani, N., Baduel, S., & Rogé, B. (2014). Visual social attention in autism spectrum disorder: Insights from eye tracking studies. *Neuroscience & Biobehavioral Reviews*, 42:279–297.
- Hammond, K. R. (1998). Ecological validity: Then and now. Retrieved from <http://www.albany.edu/cpr/brunswick/notes/essay2.html>.
- Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press.
- Joseph, R. M., Keehn, B., Connolly, C., Wolfe, J. M., & Horowitz, T. S. (2009). Why is visual search superior in autism spectrum disorder? *Developmental Science*, 12(6):1083–1096.
- Kaldy, Z., Giserman, I., Carter, A. S., & Blaser, E. (2016). The mechanisms underlying the ASD advantage in visual search. *Journal of Autism and Developmental Disorders*, 46(5):1513–1527.
- Karmiloff-Smith, A. (2012). Perspectives on the dynamic development of cognitive capacities. *Current Opinion in Neurology*, 25(2):106–111.
- Keehn, B. & Joseph, R. M. (2016). Exploring what’s missing: What do target absent trials reveal about autism search superiority? *Journal of Autism and Developmental Disorders*, pages 1–13.
- Keehn, B., Müller, R.-A., & Townsend, J. (2013). Atypical attentional networks and the emergence of autism. *Neuroscience & Biobehavioral Reviews*, 37(2): 164–183.
- Kemner, C., Ewijk, L., Engeland, H., & Hooge, I. (2008). Brief report: Eye movements during visual search tasks indicate enhanced stimulus discriminability in subjects with PDD. *Journal of Autism and Developmental Disorders*, 38 (3):553–557.
- Kingstone, A. (2009). Taking a real look at social attention. *Current Opinion in Neurobiology*, 19:52–56.

## 10. Discussion

- Kingstone, A., Smilek, D., & Eastwood, J. D. (2008). Cognitive ethology: A new approach for studying human cognition. *British Journal of Psychology*, 99(3):317–340.
- Niehorster, D. C., Cornelissen, T. H. W., Holmqvist, K., Hooge, I. T. C., & Hessels, R. S. (2017). What to expect from your remote eye tracker when participants are unrestrained. *Behavior Research Methods*.
- Oakes, L. M. (2012). Advances in eye tracking in infancy research. *Infancy*, 17(1):1–8.
- Oakes, L. M. (2010). Infancy guidelines for publishing eye-tracking data. *Infancy*, 15(1):1–5.
- Saez de Urabain, I. R., Johnson, M. H., & Smith, T. J. (2015). GraFIX: A semiautomatic approach for parsing low- and high-quality eye-tracking data. *Behavior Research Methods*, 47(1):53–72.
- Senju, A. & Johnson, M. H. (2009). Atypical eye contact in autism: models, mechanisms and development. *Neuroscience & Biobehavioral Reviews*, 33: 1204–1214.
- Shic, F., Chawarska, K., & Scassellati, B. (2008). The amorphous fixation measure revisited: With applications to autism. *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*.
- Shirama, A., Kato, N., & Kashino, M. (2016). When do individuals with autism spectrum disorder show superiority in visual search? *Autism*, pages 1–10.
- Smilek, D., Birmingham, E., Cameron, D., Bischof, W., & Kingstone, A. (2006). Cognitive ethology and exploring attention in real-world scenes. *Brain Research*, 1080:101–119.
- Song, Y., Hakoda, Y., Sanefuji, W., & Cheng, C. (2015). Can they see it? The functional field of view is narrower in individuals with autism spectrum disorder. *PLOS One*, 10(7):e0133237.
- van Berckelaer-Onnes, I. A., van de Blind, G., Anzion, P., & Werkgroep JGZ Richtlijn ASS. (2015). *JGZ-richtlijn Autismspectrumstoornissen. Signalering, begeleiding en toeleiding naar diagnostiek*. Trimbos-instituut.
- Wass, S. V., Smith, T. J., & Johnson, M. H. (2013). Parsing eye-tracking data of variable quality to provide accurate fixation duration estimates in infants and adults. *Behavior Research Methods*, 45(1):229–250.
- Wass, S. V., Forssman, L., & Leppänen, J. (2014). Robustness and precision: How data quality may influence key dependent variables in infant eye-tracker analyses. *Infancy*, 19(5):427–460.

## **11. Samenvatting in het Nederlands**

## 11.1. **Aanleiding**

Het doel van dit proefschrift is om twee mogelijke vroege voorspellers van Autisme Spectrum Stoornis (ASS) te onderzoeken. Autisme Spectrum Stoornis (ASS) is een ontwikkelingsstoornis die gekarakteriseerd wordt door 1) problemen in sociale communicatie en sociale interactie, en 2) beperkte, repetitieve gedragingen, interesses of activiteiten (American Psychiatric Association, 2013). Typische voorbeelden van problemen met sociale communicatie en interactie zijn beperkingen in het beginnen en onderhouden van een gesprek, afwijkingen in non-verbale communicatie, abnormaal oogcontact, en problemen met het ontwikkelen en onderhouden van relaties. Typische voorbeelden van repetitieve gedragingen zijn stereotype bewegingen, echolalie, inflexibiliteit in het afwijken van routines, en hypo- of hyperreactiviteit op sensorische input (bijv. beeld of geluid). Volgens schattingen is de gemiddelde leeftijd van diagnose in Nederland rond de 6 à 7 jaar (van Berckelaer-Onnes et al., 2015). Een aantal studies heeft laten zien dat de diagnose van ASS al vroeger kan dankzij nieuwe diagnostische tests (Dietz et al., 2006; Swinkels et al., 2006; Oosterling et al., 2010). Sommige studies suggereren zelfs dat de diagnose bij leeftijden onder de 3 jaar al betrouwbaar is (Moore & Goodson, 2003; Stone et al., 1999). Er zijn een aantal grootschalige projecten gestart die vroege voorspellers van ASS pogen te identificeren in het eerste jaar na de geboorte (bijv. de British Autism Study of Infant Sibling; BASIS, en European Autism Interventions; EU-AIMS). Binnen deze projecten worden baby's gevolgd die al een oudere broer of zus hebben met een ASS diagnose. Aangezien ASS voor een deel erfelijk is, hebben deze baby's een groter risico op een diagnose van ASS dan baby's die een oudere broer of zus zonder diagnose hebben. De belangrijkste redenen om mogelijke vroege voorspellers van ASS te onderzoeken zijn 1) een beter beeld krijgen van de kenmerken van ASS tijdens de ontwikkeling (bijv. Karmiloff-Smith, 2012), en 2) vroegere diagnose en interventie mogelijk maken.

In dit proefschrift worden twee mogelijke vroege voorspellers van ASS op basis van kijkgedrag bestudeerd: 1) superieure visuele zoekprestatie en 2)

kijkgedrag tijdens het bekijken van gezichten. Eerder onderzoek heeft aangetoond dat volwassenen met ASS hierop te onderscheiden zijn van typisch ontwikkelde (TO) volwassenen. Individuen met ASS vertonen superieure zoekprestatie vergeleken met TO individuen, in het bijzonder in moeilijke zoektaken waar het zoekdoel lijkt op de niet-doelen. In het onderzoeksveld gezichtswaarneming is gerapporteerd dat individuen met ASS minder lang naar de ogen kijken op een gezicht dan TO individuen. Hier is echter geen consensus over in de literatuur. De twee mogelijke vroege voorspellers zijn gericht op de verschillende diagnostische criteria van ASS. De eerste mogelijk voorspeller, superieure visuele zoekprestatie, is typerend voor de visuele eigenaardigheden van ASS (zie bijv. Simmons et al., 2009). De tweede mogelijke voorspeller, kijkgedrag tijdens het bekijken van gezichten, is daarentegen typerend voor de sociale problematiek in ASS.

Al in het begin van dit promotietraject werd duidelijk dat het onderzoek naar vroege voorspellers van ASS niet uit te voeren is met de huidige technieken. Het onderzoek naar beide vroege voorspellers van ASS is afhankelijk van technologie om kijkgedrag te bestuderen. Dit gebeurt met zogenaamde 'eye trackers'. Vrijwel alle eye trackers op de markt zijn ontworpen voor en getest op volwassen proefpersonen. Het is maar de vraag of het gebruik van deze eye trackers bij baby's bruikbare data oplevert. In dit proefschrift wordt daarom eerst de beschikbare methodologie voor het bestuderen van kijkgedrag bij baby's kritisch onderzocht. Voor het onderzoek naar vroege voorspellers van ASS is niet alleen adequate methodologie vereist, maar ook kennis over typische ontwikkeling. Over één van de mogelijke vroege voorspellers, te weten visuele zoekprestatie, was er nog weinig bekend in typische ontwikkeling. Tenslotte bleek er binnen de gezichtswaarneming geen consensus over de vraag of individuen met ASS minder naar de ogen kijken dan TO individuen. Omdat veel van dit onderzoek met statische representaties (foto's van gezichten) is uitgevoerd, onderzoeken we hier kijkgedrag tijdens het bekijken van gezichten in sociale interactie (d.w.z. gezichten van personen die echt aanwezig zijn). Om dit mogelijk te maken, moest er nieuwe methodologie ontwikkeld worden. Hierdoor is een groot deel van dit proefschrift gericht op ontwikkelingen die toekomstig onderzoek naar

## 11. *Samenvatting in het Nederlands*

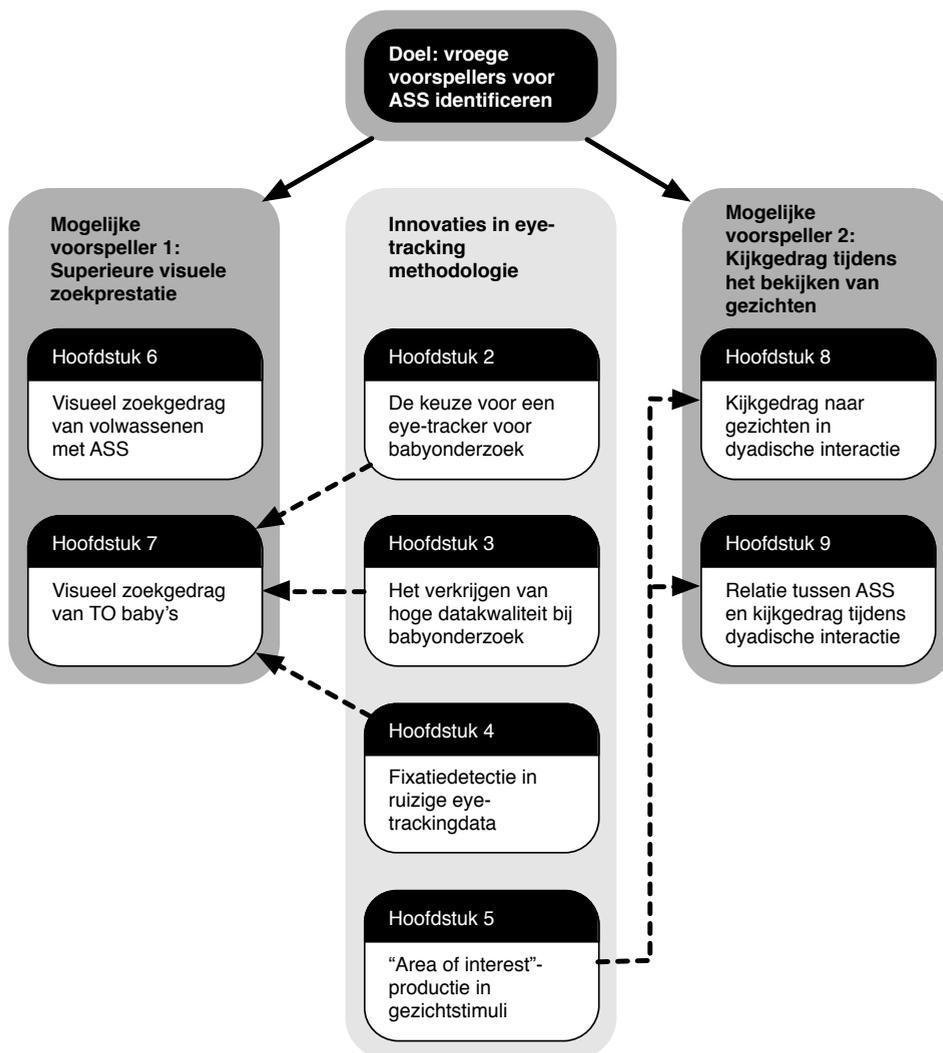
vroege voorspellers van ASS mogelijk maakt.

In deze samenvatting worden achtereenvolgend de volgende onderwerpen behandeld. Ten eerste wordt de methodiek besproken, die gebruikt wordt om kijkgedrag bij baby's te bestuderen. Ten tweede wordt visuele zoekprestatie als mogelijke voorspeller van ASS besproken. De bestaande literatuur over visueel zoekgedrag in ASS kan als volgt worden samengevat: individuen met ASS zijn sneller in het lokaliseren van een zoekdoel dan TO individuen zonder dat ze minder accuraat zijn. Dit is gerapporteerd voor verschillende leeftijdsgroepen, van peuterleeftijd tot volwassenen. In dit proefschrift worden twee studies beschreven naar visueel zoekgedrag: 1) de bovengrens van de superieure zoekprestatie van individuen met ASS en 2) visueel zoekgedrag van baby's. Ten derde wordt een nieuwe aanpak besproken voor onderzoek naar kijkgedrag tijdens het bekijken van gezichten. In Figuur 11.1 staat een schematisch overzicht van de hoofdstukken met hun onderliggende relaties.

### **11.2. Methodiek voor het bestuderen van kijkgedrag**

Voor een eye-trackingstudie worden er een aantal stappen doorlopen.

1. Er wordt een geschikte eye tracker uitgekozen. De participant wordt zorgvuldig voor de eye tracker gepositioneerd.
2. Het signaal van de eye tracker wordt gekalibreerd en gevalideerd door de proefpersoon naar een aantal vaste punten op het beeldscherm van de eye tracker te laten kijken.
3. Visuele stimuli worden aangeboden op het scherm, en eye-trackingdata wordt verzameld.
4. Om uitspraken te doen over waar een participant gekeken heeft op een visuele stimulus, wordt de eye-trackingdata geclassificeerd in oogbewegingen en fixaties (periodes waarin het oog niet beweegt). Hierna worden deze fixaties gekoppeld aan delen van de visuele stimulus.



Figuur 11.1.: Schematisch overzicht van de hoofdstukken van dit proefschrift. Pijlen geven aan welke studies een randvoorwaarde zijn voor andere studies.

## 11. Samenvatting in het Nederlands

5. Nadat fixaties aan de visuele stimulus gekoppeld zijn, worden uitkomstmaten uitgerekend: bijvoorbeeld het aantal keer dat er in een interessegebied (*Area of Interest*, kortweg AOI) wordt gekeken, of de gemiddelde tijd dat er in een AOI wordt gekeken. Statistische analyse kan vervolgens op deze maten uitgevoerd worden.

Hoewel al deze stappen goed onderzocht zijn voor onderzoek met volwassen participanten (Holmqvist et al., 2011), is er weinig bekend over deze stappen als de proefpersonen baby's of jonge kinderen zijn. De voordelen van het gebruik van eye tracking bij babyonderzoek zijn volgens verscheidene onderzoekers voor de hand liggend (Aslin & McMurray, 2004; Oakes, 2012): Met eye tracking is het immers mogelijk om visuele verwerking, interesses, en voorkeur, van niet-verbale proefpersonen te onderzoeken. De nadelen van deze techniek zijn echter niet vaak genoemd (Wass et al., 2013, 2014). Zo is de kwaliteit van eye-trackingdata verzameld bij baby's vaak lager dan dat verzameld met volwassenen als participanten. Om deze problematiek aan te pakken moet een onderzoeker iets van signaalverwerking weten. Deze kennis was bij eerdere methodes om kijkgedrag van baby's te onderzoeken (bijvoorbeeld observationele technieken) niet nodig. In de bovenstaande stappen moet een grote hoeveelheid keuzes gemaakt worden, die allen een effect kunnen hebben op de uitkomstmaten van het onderzoek. Meer kennis over hoe eye-trackingdata van hoge kwaliteit bij baby's en kinderen verzameld kan worden is daarom hard nodig. In dit proefschrift zijn alle bovenstaande stappen onderzocht.

In **Hoofdstuk 2** werd de keuze voor een eye tracker voor gebruik met moeilijke proefpersoongroepen onderzocht (stap 1 in bovenstaand overzicht). Acht verschillende eye trackers werden onderzocht met betrekking tot 1) hoe een eye tracker omgaat met het terugvinden van de blikrichting nadat de proefpersoon van het scherm afgekeken heeft (dit is niet triviaal) en 2) hoe eye trackers presteren tijdens niet-optimale hoofdorïentaties. Onder niet-optimale hoofdorïentaties vallen alle posities waarin de participant niet rechtop voor de eye tracker zit, bijvoorbeeld als iemand zijn hoofd scheef houdt. De proefpersonen waren volwassenen, die oog- en hoofdbewe-

gingen van baby's nabootsten. De verschillende eye trackers verschilden in de hoeveelheid gemiste metingen (dataverlies) tijdens niet-optimale hoofdoriëntaties. Bovendien verschilden eye trackers in de systematische fout – het verschil tussen de echte kijkpositie van een proefpersoon, en de gerapporteerde kijkpositie door de eye tracker – die ze rapporteerden als er wel een kijkpositie gerapporteerd werd. Omdat fabrikanten vaak datakwaliteit rapporteren gemeten in optimale omstandigheden, raden we oogbewegingsonderzoekers aan om hun eye trackers ook te testen onder niet-optimale omstandigheden. Hierdoor krijgen onderzoekers een beeld van de datakwaliteit onder realistische omstandigheden.

Nadat er een keuze voor een eye tracker is gemaakt, is de volgende stap om ervoor te zorgen dat er goede data verzameld wordt (stap 2 en 3 in bovenstaand overzicht). Eye-trackingdata is van hoge kwaliteit als deze een kleine systematische fout, een kleine variabele fout, en geen tot weinig dataverlies bevat. In **Hoofdstuk 3** onderzochten we hoe de kwaliteit van eye-trackingdata verzameld bij baby's afhangt van kleur van de ogen, positionering (of een baby in een kinderstoel zit, of op de schoot van de ouder), aantal calibraties, en de tijd sinds het kind gegeten en geslapen heeft. We rapporteren dat data verzameld met kinderen met blauwe ogen van lagere kwaliteit is dan data van kinderen met bruine ogen. Naarmate het experiment vorderde, nam het dataverlies toe. Tenslotte rapporteren we dat de positionering van de baby de systematische fout in de eye-trackingdata lijkt te beïnvloeden: eye-tracking data van baby's die in een kinderstoel zaten bevatte een grotere systematische fout dan eye-tracking data van baby's die in een babyzitje zaten op de schoot van de ouder. Als een onderzoeker twee groepen wil vergelijken op een uitkomstmaat (bijvoorbeeld aantal fixaties), is het van belang dat deze groepen niet verschillen op bovenstaande voorspellers van datakwaliteit. Als dit wel zo is, kunnen verschillen tussen de groepen namelijk aan datakwaliteit toe te schrijven zijn, en niet aan een daadwerkelijk verschil tussen de groepen.

In **Hoofdstuk 4** introduceerden we een nieuw fixatiedetectiealgoritme, dat ontwikkeld is om automatische data-analyse van eye-trackingdata van

## 11. Samenvatting in het Nederlands

verschillende ruis- en dataverliesniveau's mogelijk te maken (stap 4 in bovenstaand overzicht). Het algoritme “*Identification by 2-means clustering*”, ofwel I2MC, werd getest tegen 7 andere moderne algoritmes. We rapporteren dat de uitkomstmaten (aantal fixaties en gemiddelde fixatieduur) van het I2MC-algoritme het minst veranderen als het ruis- en dataverliesniveau toeneemt in vergelijking met de 7 andere algoritmes. Deze bevinding is van cruciaal belang voor onderzoek naar vroege voorspellers van ASS. Zo lieten Shic et al. (2008) zien dat de parameterkeuze voor een fixatiedetectiealgoritme kan bepalen wat de richting is van een verschil tussen ASS and TO kinderen, omdat de datakwaliteit tussen deze twee groepen niet gelijk is. Doordat het I2MC-algoritme in data van hoge en (tot op zekere hoogte) lage kwaliteit dezelfde uitkomst geeft, hebben datakwaliteitsverschillen tussen twee groepen minder tot geen invloed op de conclusies. Bovendien werkt het I2MC-algoritme volledig automatisch, waardoor handmatige correctie van fixatiedetectie niet meer nodig is (Saez de Urabain et al., 2015).

In **Hoofdstuk 5** onderzochten we verschillende methoden om AOIs te produceren voor gezichtsstimuli (stap 4 en 5 in bovenstaand overzicht). We rapporteren dat AOI-gebaseerde uitkomstmaten niet aanzienlijk verschillen tussen verschillende AOI-productiemethoden. Voor AOIs geproduceerd met de Voronoi-methode of de “Limited-Radius Voronoi Tessellation” (LRVT) methode hoeft de onderzoeker echter het minst aantal subjectieve keuzes te maken. Met deze methodes is het namelijk niet nodig de grootte of vorm van een AOI expliciet te definiëren. Bovendien kunnen de Voronoi- en LRVT-methode automatisch geïmplementeerd worden met een computerscript. Deze automatische implementatie kan in vergelijking met handmatige AOI-productie leiden tot aanzienlijke afname van de tijd die een onderzoeker besteedt aan het produceren van AOIs. Tenslotte concluderen we dat uitkomstmaten van grote AOIs in gezichtsstimuli het meest robuust zijn tegen toenames in variabele fout (ofwel ruis). Als de participanten baby's of jonge kinderen zijn kan het zijn dat er eye-trackingdata van lage kwaliteit verzameld wordt. In dat geval raden we het gebruik van grote AOIs aan als de stimuli uit gezichten bestaat, of maar weinig elementen bevat.

Met behulp van de hier beschreven studies zijn de keuzes die een onderzoeker maakt in bovenstaande stappen beter te beargumenteren. Dit betreft de keuze voor een eye tracker, het verkrijgen van goede data, het analyseren van data van lage kwaliteit, en het produceren van AOIs. Dit betekent echter niet dat er nu een standaard is voor het onderzoek naar kijkgedrag bij baby's. Daar zijn verscheidene redenen voor. Ten eerste werken fabrikanten constant aan nieuwe eye trackers, die beter presteren in optimale en niet-optimale omstandigheden ten opzichte van voorgaande eye trackers. Om een overzicht te behouden van geschikte eye trackers voor babyonderzoek, is het belangrijk dat deze nieuwe eye trackers getest worden onder niet-optimale omstandigheden. Als opvolging van de studie beschreven in **Hoofdstuk 2** hebben wij recentelijk een nieuwe vergelijking van eye trackers ondernomen (Niehorster et al., 2017). Ten tweede is de inhoud van de hoofdstukken over data-analyse in dit proefschrift beperkt tot het detecteren van fixaties en het produceren van AOIs voor gezichtsstimuli. Het ligt voor de hand dat babyonderzoek zich niet beperkt tot deze situaties en variabelen. Het is belangrijk om ook analysemethoden voor eye-trackingdata van lage kwaliteit voor andere stimuli te ontwikkelen, of voor verschillende oogbewegingen.

### 11.3. Visueel zoekgedrag in ASS

Een groot aantal studies heeft gerapporteerd dat individuen met ASS superieure zoekprestatie vertonen in vergelijking met TO controles. In **Hoofdstuk 6** beschrijven we een visueel zoekexperiment met een ASS- en controlegroep. Tegen onze verwachting in was de zoekprestatie van de ASS-groep niet superieur aan de prestatie van de controlegroep. In een eerdere studie uit ons laboratorium observeerden we namelijk wel superieure zoekprestatie van de ASS groep met een vergelijkbare zoektaak (Kemner et al., 2008). Het bleek dat er in het experiment in **Hoofdstuk 6** zoekschermen gebruikt waren waarin het verschil tussen het zoekdoel (een verticale lijn) en de niet-doelen (gedraaide lijnen) kleiner was dan in het experiment van Kemner et al. (2008). Dit betekent dat het zoekdoel in ons experiment

## 11. *Samenvatting in het Nederlands*

lastiger te lokaliseren was dan in de studie van Kemner et al. (2008). Op basis van deze informatie werd een tweede experiment afgenomen waarin het experiment uit Kemner et al. (2008) herhaald werd. In dit tweede experiment werd wel superieure zoekprestatie van de ASS-groep geobserveerd ten opzichte van de controlegroep. Op basis van deze experimenten formuleerden we een vermoeden dat perceptuele belasting bepaalt of superieure zoekprestatie van een ASS-groep geobserveerd wordt. Perceptuele belasting is de hoeveelheid, en de complexiteit van de, informatie die er op het perceptuele systeem wordt gelegd. Zolang de perceptuele belasting capaciteit over laat voor het verwerken van taak-(ir)relevante informatie, zal er superieure zoekprestatie geobserveerd worden voor individuen met ASS ten opzichte van controles. Als de perceptuele belasting dusdanig hoog is dat er geen capaciteit over is voor het verwerken van taak-(ir)relevante informatie, dan zal er geen superieure zoekprestatie voor individuen met ASS geobserveerd worden.

Superieure zoekprestatie van individuen met ASS wordt in **Hoofdstuk 6** verklaard aan de hand van de perceptuele belastingtheorie. Er zijn verscheidene andere theoriën die de superieur zoekprestatie van individuen met ASS gepoogd hebben te verklaren. Deze theoriën hebben echter geen testbare hypothesen opgeleverd, noch superieure zoekprestatie in ASS voldoende weten te verklaren (Kaldy et al., 2016; Keehn et al., 2013). Hierom stellen we voor de verschillen in het zoekgedrag tussen individuen met ASS en TO individuen beter in kaart te brengen. Als we het zoekgedrag van individuen met ASS beter begrijpen, kunnen we wellicht betere modellen genereren. Hiervoor moeten we het kijkgedrag tijdens visuele zoektaken goed beschrijven. Een eerste studie heeft deze aanpak al toegepast (zie bijv. Keehn & Joseph, 2016).

Om visueel zoekgedrag als een vroege voorspeller voor ASS te onderzoeken, is kennis vereist over zoekgedrag van TO baby's. In **Hoofdstuk 7** onderzochten we het zoekgedrag van TO baby's. Ten eerste onderzochten we of baby's van 10 maanden spontaan zoekgedrag vertonen. Aangezien baby's geen instructies kunnen volgen, onderzochten we of baby's zoekge-

drag vertoonden wanneer ze blootgesteld werden aan stimuli die normaal gesproken gebruikt worden om zoekgedrag bij volwassenen te onderzoeken. Ten tweede karakteriseerden we het zoekgedrag van TO baby's en beschreven de gelijkenissen met zoekgedrag van volwassenen. Als men uitspraken wil doen op een individueel niveau over (a)typisch zoekgedrag is het tevens belangrijk dat maten van zoekgedrag betrouwbaar zijn. Om uitspraak te doen over de betrouwbaarheid van uitkomstmaten van zoekgedrag berekenen we de test-hertestbetrouwbaarheid van deze maten.

Baby's van 10 maanden oud vertoonden spontaan zoekgedrag. Dit concludeerden we op basis van het feit dat zoekprestatie afhankelijk was van het verschil tussen het zoekdoel en de niet-doelen. Als het verschil tussen het zoekdoel en niet-doelen klein is, is het zoekdoel lastiger te lokaliseren en zal de gemiddelde zoektijd langer zijn. Zoekprestatie was bovendien beter dan wat we op basis van kans konden verwachten. Het zoekgedrag van baby's liet gelijkenissen zien met zoekgedrag van volwassenen. Zo was de eerste fixatie tijdens een trial, die gezien kan worden als de latentie totdat de eerste oogbeweging gemaakt wordt, langer dan de opvolgende fixaties. Daarbij gingen oogbewegingen vaak dezelfde richting op als voorgaande oogbewegingen. Dit gebeurde dan met dezelfde amplitude. Als oogbewegingen niet dezelfde richting op gingen, gingen ze vaak terug naar waar er net gekeken was. De test-hertest betrouwbaarheid was het hoogst voor fixatieduur en latentie om een oogbeweging te maken en minder voor zoekprestatie. Deze bevindingen betekenen dat uitspraken over (a)typisch zoekgedrag op individueel niveau vooralsnog onbetrouwbaar zijn.

Hoewel we hebben laten zien dat baby's van 10 maanden spontaan zoekgedrag vertonen, dat tevens gelijkenissen met volwassen zoekgedrag laat zien, blijven er een paar vragen over. Ten eerste vragen we ons af hoe zoekgedrag ontwikkelt vanaf de geboorte tot de peutertijd? Vertonen baby's ook al spontaan zoekgedrag voordat ze 10 maanden oud zijn? Deze vragen zijn van belang om zoekgedrag in baby's te kunnen bestempelen als atypisch.

## 11.4. **Kijkgedrag in ASS tijdens het bekijken van gezichten**

Zoals eerder beschreven, bleek er binnen de gezichtswaarneming geen consensus te zijn over de vraag of individuen met ASS minder naar de ogen kijken dan TO individuen. Zo concludeerden de auteurs van een recent literatuuroverzicht dat de hypothese dat individuen met ASS meer naar de mond kijken en minder naar de ogen kijken dan TO controles, weinig ondersteuning vindt (Guillon et al., 2014). De auteurs van een eerder literatuuroverzicht concludeerden dat de verminderde kijktijd naar de ogen van individuen met ASS het meest prominent is onder hoge cognitieve belasting (Senju & Johnson, 2009). Het lijkt wellicht verrassend dat er geen consensus bestaat over abnormaal kijkgedrag tijdens het bekijken van gezichten, aangezien ASS met grote beperkingen in sociale communicatie en interactie gepaard gaat. Als men de studies er op naloopt, blijkt dit echter minder verrassend. In veel studies waarin het kijkgedrag tijdens het bekijken van gezichten onderzocht werd, werden namelijk niet-responsieve stimuli (bijvoorbeeld foto's of video's van gezichten) gebruikt. Het kijkgedrag naar een niet-responsief gezicht wordt vervolgens beschouwd als model voor sociale interactie. In het dagelijks leven behoren gezichten echter vaak tot een persoon waarmee geïnteracteed kan worden. Er kan dan ook niet zomaar gegeneraliseerd worden van kijkgedrag naar een foto van een gezicht naar kijkgedrag met echte mensen. Dit bezwaar heeft zelfs een beweging doen ontstaan, die voorstelt om gedrag eerst in natuurlijke situaties te bestuderen en dan pas in het laboratorium (Smilek et al., 2006; Kingstone et al., 2008; Kingstone, 2009).

Om kijkgedrag van twee personen die interacteren met hoge spatiële en temporele resolutie te kunnen bestuderen, ontwikkelden we een interactieopstelling. In dit proefschrift zijn twee studies beschreven waar gebruik is gemaakt van deze interactieopstelling. In **Hoofdstuk 8** onderzochten we of kijkgedrag tijdens het bekijken van gezichten in dyadische interactie (interactie van twee personen) vergelijkbaar was met kijkgedrag naar afbeeldingen of video's van gezichten. Specifiek onderzochten we of parti-

#### 11.4. Kijkgedrag in ASS tijdens het bekijken van gezichten

cipanten een voorkeur hadden voor het bekijken van de ogen van de ander. We rapporteren dat er in dyadische interactie een voorkeur was voor het kijken naar de oogregio van een ander, net zoals dat gerapporteerd is voor foto's en video's van anderen. Tevens lijkt het erop dat kijkgedrag tijdens het bekijken van gezichten gekoppeld was tussen twee interactiepartners. De totale tijd dat één partner naar de ogen keek bleek namelijk een goede voorspeller voor de tijd dat de ander naar de ogen keek. Wanneer we het kijkgedrag van één interactiepartner expliciet manipuleerden door deze via een oortje kijkinstructies te geven, observeerden we echter niet dat het kijkgedrag van de ander hiervan afhankelijk was. Als er een koppeling bestaat tussen het kijkgedrag van twee interactiepartners, dan lijkt deze niet gemakkelijk te manipuleren.

In **Hoofdstuk 9** onderzochten we of kijkgedrag tijdens het bekijken van gezichten in dyadische interactie gerelateerd was aan sub-klinische psychopathologie. Specifiek onderzochten we of het kijkgedrag gerelateerd was aan karaktertrekken van ASS en Sociale Angst Stoornis (SAS). Uit eerder onderzoek naar kijkgedrag van patiënten met ASS en SAS bleek dat deze patiënten minder naar de ogen keken dan TO controles. Hoewel dit onderzoek met foto's van gezichten werd uitgevoerd, verwachtten we ook in dyadische interactie dat participanten die hoog scoorden op karaktertrekken van ASS en SAS, minder naar de oogregio van de ander zouden kijken. We rapporteren dat participanten die hoog scoorden op karaktertrekken van ASS en SAS inderdaad minder naar de ogen van de ander keken dan participanten die laag op deze karaktertrekken scoorden. Wanneer we participanten als paren beschouwden, observeerden we relaties tussen kijkgedrag en karaktertrekken van ASS en SAS. Paren die hoog scoorden op ASS-karaktertrekken, bevonden zich langer in de situatie dat slechts één van de twee naar de ogen van de ander keek, dan paren die laag scoorden op ASS-karaktertrekken. Deze hoog-scorende paren bevonden zich tevens korter in de situatie waarin beide participanten elkaar in de ogen keken, vergeleken met laag-scorende paren. Paren die hoog scoorden op SAS-karaktertrekken bevonden zich vaker in de situatie dat slechts één van de participanten de ander in de ogen keek dan paren die laag scoorden op SAS-

## 11. *Samenvatting in het Nederlands*

karaktertrekken. Als hoog-scorende paren zich in deze situatie bevonden, was dat echter van kortere gemiddelde duur, dan laag-scorende paren. Kijkgedrag van twee interacterende personen blijkt dus verschillend gerelateerd aan ASS- en SAS-karaktertrekken.

**Hoofdstukken 8-9** laten belangrijke overeenkomsten zien in kijkgedrag tijdens het bekijken van gezichten tussen responsieve stimuli (personen die kunnen interacteren) en niet-responsieve representaties (foto's en video's). Deze overeenkomsten zijn 1) er is een voorkeur voor het bekijken van de ogen van een ander en 2) personen die hoog scoren op ASS- en SAS-karaktertrekken kijken minder naar de ogen dan personen die laag scoren op deze karaktertrekken. De interactieopstelling heeft echter ook nieuwe perspectieven op gezichtswaarneming mogelijk gemaakt. Zo rapporteerden we dat paar-gebaseerde maten van kijkgedrag verschillend gerelateerd zijn aan ASS- en SAS-karaktertrekken. De interactieopstelling maakt het ook mogelijk nieuwe vragen te beantwoorden, zoals: In hoeverre wordt het kijkgedrag van een persoon bepaald door gedrag van de ander, bijvoorbeeld als iemand de ander niet of wel aankijkt? Hoe is kijkgedrag afhankelijk van een vertraging in de videoverbinding? Wanneer hebben personen door dat de ander niet meer 'live' is?

Eén cruciale vraag is of de interactieopstelling de beste manier is om kijkgedrag tijdens het bekijken van gezichten als vroege voorspeller van ASS te onderzoeken. Recent onderzoek doet dit sterk vermoeden. Edmunds et al. (2017) waren geïnteresseerd in het meten van oogcontact tussen kinderen met ASS en hun sociale partners (bijv. ouders). Ze rapporteren dat het onmogelijk is om dit met video-observaties of mobiele eye trackers (bijv. brillen die kijkgedrag meten) te meten. De spatiële resolutie van deze technieken is namelijk niet hoog genoeg om gedetailleerdere uitspraken te doen dan "kijkt naar gezicht" en "kijkt niet naar gezicht". Edmunds et al. (2017) lossen dit probleem op door een camerabril op het gezicht van de sociale partner te plaatsen. Alleen door het videobeeld van deze camera te gebruiken, konden ze verschillen in oogbewegingsgedrag tussen kinderen met en zonder ASS oppikken. Onze interactieopstelling maakt het moge-

lijk om met dusdanig hoge resolutie te meten, dat het objectief meten van oogcontact mogelijk is tijdens interactie tussen kinderen met ASS en hun sociale partners. Daarnaast is het mogelijk om gedetailleerdere analyses van kijkgedrag in interactie uit te voeren, wat nieuwe perspectieven opent voor het onderzoek naar (a)typisch kijkgedrag tijdens gezichtswaarneming.

## 11.5. Conclusies

In dit proefschrift zijn twee mogelijke vroege voorspellers voor ASS onderzocht: superieure visuele zoekprestatie, en kijkgedrag tijdens het bekijken van gezichten. Al in het begin van dit promotietraject werd duidelijk dat het onderzoek naar vroege voorspellers van ASS niet uit te voeren is met de huidige technieken. Hierom is eerst de beschikbare methodologie voor het besturen van babykijkgedrag kritisch onderzocht. Ten tweede bleek dat er nog weinig kennis was over typisch zoekgedrag van baby's. Tenslotte bleek er binnen de gezichtswaarneming geen consensus over de vraag of individuen met ASS minder naar de ogen kijken dan TO individuen. Omdat dit mogelijk kan liggen aan het feit dat veel onderzoek niet-responsieve representaties van gezichten heeft gebruikt, ontwikkelden we een opstelling om kijkgedrag in sociale interactie te onderzoeken.

Dankzij de studies omschreven in **Hoofdstuk 2-5** kan een oogbewegingsonderzoeker beter beargumenteren welke keuzes er gemaakt worden: van de keuze voor een eye tracker, het verkrijgen van goede data, het detecteren van fixaties in ruizige data, en de keuze voor AOI-productiemethode. Deze studies waren nodig 1) om visueel zoekgedrag van typische-ontwikkende baby's te onderzoeken en 2) om automatische data-analyse van kijkgedrag in dyadische interactie mogelijk te maken. In **Hoofdstuk 6** werd visueel zoekgedrag van volwassenen met ASS onderzocht, en de bevindingen zijn in context geplaatst van de huidige theoriën over zoekgedrag in ASS. De huidige theoriën geven geen duidelijke beschrijving van zoekprestatie in ASS, noch hebben deze testbare hypothesen gegenereerd. Er wordt daarom voorgesteld het kijkgedrag van individuen met ASS tijdens visuele zoektaken gedetailleerder in kaart te brengen. Dit maakt het wellicht mogelijk

## *11. Samenvatting in het Nederlands*

betere modellen van visuele zoekprestatie in ASS te genereren. In **Hoofdstuk 7** werd visueel zoekgedrag van typisch-ontwikkellende baby's onderzocht. Baby's vertoonden spontaan zoekgedrag dat gelijkenissen vertoonde met volwassen zoekgedrag. In **Hoofdstuk 8** werd een nieuwe methode voor het onderzoeken van kijkgedrag tijdens het bekijken van gezichten geïntroduceerd en in **Hoofdstuk 9** werd het kijkgedrag tijdens dyadische interactie gerelateerd aan ASS- en SAS-karaktertrekken. Kijkgedrag tijdens het bekijken van gezichten in dyadische interactie vertoonde belangrijke overeenkomsten met kijkgedrag naar niet-responsieve representaties van gezichten. Bovendien identificeerden we relaties tussen het kijkgedrag van twee interactiepartners en ASS- en SAS-karaktertrekken.

Concluderend worden er in dit proefschrift belangrijke methodologische innovaties beschreven die cruciaal zijn om typisch en atypisch kijkgedrag te beschrijven en te begrijpen. Dit geldt niet alleen voor kijkgedrag tijdens de vroege ontwikkeling, maar ook bij volwassenen. Tevens zijn deze innovaties van groot belang voor ons begrip van de ontwikkeling van visueel zoekgedrag van de babytijd tot volwassenheid. Bovendien introduceerden we een nieuwe opstelling die het mogelijk maakt kijkgedrag tijdens dyadische interactie te onderzoeken. Met deze interactieopstelling hebben we al intrigerende relaties laten zien tussen kijkgedrag in sociale interactie en sub-klinische psychopathologie. Onze interactieopstelling maakt nieuwe perspectieven mogelijk op het onderzoek naar (a)typisch kijkgedrag in sociale interactie, en kan ons helpen de sociale problematiek in ASS beter te begrijpen.

## Referenties

- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders*. American Psychiatric Association, Washington, DC, 5th edition.
- Aslin, R. N. & McMurray, B. (2004). Automated corneal-reflection eye tracking in infancy: Methodological developments and applications to cognition. *Infancy*, 6(2):155–163.
- Dietz, C., Swinkels, S., van Daalen, E., van Engeland, H., & Buitelaar, J. K. (2006). Screening for autistic spectrum disorder in children aged 14–15 months. II: Population screening with the early screening of autistic traits questionnaire (ESAT). Design and general findings. *Journal of Autism and Developmental Disorders*, 36(6):713–722.
- Edmunds, S. R., Rozga, A., Li, Y., Karp, E. A., Ibanez, L. V., Rehg, J. M., & Stone, W. L. (2017). Brief report: Using a point-of-view camera to measure eye gaze in young children with autism spectrum disorder during naturalistic social interactions: A pilot study. *Journal of Autism and Developmental Disorders*, pages 1–7.
- Guillon, Q., Hadjikhani, N., Baduel, S., & Rogé, B. (2014). Visual social attention in autism spectrum disorder: Insights from eye tracking studies. *Neuroscience & Biobehavioral Reviews*, 42:279–297.
- Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press.
- Kaldy, Z., Giserman, I., Carter, A. S., & Blaser, E. (2016). The mechanisms underlying the ASD advantage in visual search. *Journal of Autism and Developmental Disorders*, 46(5):1513–1527.
- Karmiloff-Smith, A. (2012). Perspectives on the dynamic development of cognitive capacities. *Current Opinion in Neurology*, 25(2):106–111.
- Keehn, B. & Joseph, R. M. (2016). Exploring what’s missing: What do target absent trials reveal about autism search superiority? *Journal of Autism and Developmental Disorders*, pages 1–13.
- Keehn, B., Müller, R.-A., & Townsend, J. (2013). Atypical attentional networks and the emergence of autism. *Neuroscience & Biobehavioral Reviews*, 37(2): 164–183.
- Kemner, C., Ewijk, L., Engeland, H., & Hooge, I. (2008). Brief report: Eye movements during visual search tasks indicate enhanced stimulus discriminability

## 11. Samenvatting in het Nederlands

- in subjects with PDD. *Journal of Autism and Developmental Disorders*, 38 (3):553–557.
- Kingstone, A. (2009). Taking a real look at social attention. *Current Opinion in Neurobiology*, 19:52–56.
- Kingstone, A., Smilek, D., & Eastwood, J. D. (2008). Cognitive ethology: A new approach for studying human cognition. *British Journal of Psychology*, 99 (3):317–340.
- Moore, V. & Goodson, S. (2003). How well does early diagnosis of autism stand the test of time? *Autism*, 7(1):47–63.
- Niehorster, D. C., Cornelissen, T. H. W., Holmqvist, K., Hooge, I. T. C., & Hessels, R. S. (2017). What to expect from your remote eye tracker when participants are unrestrained. *Behavior Research Methods*.
- Oakes, L. M. (2012). Advances in eye tracking in infancy research. *Infancy*, 17 (1):1–8.
- Oosterling, I. J., Wensing, M., Swinkels, S. H., van der Gaag, R. J., Visser, J. C., Woudenbergh, T., Minderaa, R., Steenhuis, M.-P., & Buitelaar, J. K. (2010). Advancing early detection of autism spectrum disorder by applying an integrated two-stage screening approach. *Journal of Child Psychology and Psychiatry*, 51(3):250–258.
- Saez de Urabain, I. R., Johnson, M. H., & Smith, T. J. (2015). GraFIX: A semiautomatic approach for parsing low- and high-quality eye-tracking data. *Behavior Research Methods*, 47(1):53–72.
- Senju, A. & Johnson, M. H. (2009). Atypical eye contact in autism: models, mechanisms and development. *Neuroscience & Biobehavioral Reviews*, 33: 1204–1214.
- Shic, F., Chawarska, K., & Scassellati, B. (2008). The amorphous fixation measure revisited: With applications to autism. *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*.
- Simmons, D. R., Robertson, A. E., McKay, L. S., Toal, E., McAleer, P., & Pollick, F. E. (2009). Vision in autism spectrum disorders. *Vision Research*, 49(22): 2705–2739.
- Smilek, D., Birmingham, E., Cameron, D., Bischof, W., & Kingstone, A. (2006). Cognitive ethology and exploring attention in real-world scenes. *Brain Research*, 1080:101–119.
- Stone, W. L., Lee, E. B., Ashford, L., Brissie, J., Hepburn, S. L., Coonrod, E. E., & Weiss, B. H. (1999). Can autism be diagnosed accurately in children under 3 years? *Journal of Child Psychology and Psychiatry*, 40(2):219–226.

- Swinkels, S. H. N., Dietz, C., van Daalen, E., Kerkhof, I. H. G. M., van Engeland, H., & Buitelaar, J. K. (2006). Screening for autistic spectrum in children aged 14 to 15 months. I: The development of the early screening of autistic traits questionnaire (ESAT). *Journal of Autism and Developmental Disorders*, 36(6):723–732.
- van Berckelaer-Onnes, I. A., van de Blind, G., Anzion, P., & Werkgroep JGZ Richtlijn ASS. (2015). *JGZ-richtlijn Autismespectrumstoornissen. Signalering, begeleiding en toeleiding naar diagnostiek*. Trimbos-instituut.
- Wass, S. V., Smith, T. J., & Johnson, M. H. (2013). Parsing eye-tracking data of variable quality to provide accurate fixation duration estimates in infants and adults. *Behavior Research Methods*, 45(1):229–250.
- Wass, S. V., Forssman, L., & Leppänen, J. (2014). Robustness and precision: How data quality may influence key dependent variables in infant eye-tracker analyses. *Infancy*, 19(5):427–460.



## **12. Acknowledgements**

## 12. Acknowledgements

This dissertation would never have been possible without a large number of people. Given the multilingual background of those people, these acknowledgements are too.

Ten eerste, Chantal, ik weet niet of je je ons eerste contact nog kan herinneren. Ik in ieder geval wel. Ik was zo'n 4 maanden afgestudeerd, en was net vertrokken op een vakantie van een maand. Al rijdend door de Schotse hooglanden kreeg ik een mail van je. Of ik interesse had me in een visuele zoekproef bij volwassen autisten te verdiepen, voor een periode van een jaar. Interesse had ik wel, maar (in alle onwetendheid over wie je was) vroeg ik of het oké was als ik over een maand contact opnam. "*Prima!*", mailde je. Zo zaten we een maand later aan tafel, samen met Ignace. Ik vertelde je dat ik nog wel een andere sollicitatie had lopen, en wilde nadenken over wat ik precies wilde doen. Weer geen probleem, zei je. Een dag of 2 later besloot ik de baan aan te nemen, en hoewel ik de consequenties van de andere keuze niet kon overzien, was ik ervan overtuigd dat het de beste beslissing was die ik kon maken. Al na een paar maanden sprak je de intentie uit om mij nog 3 jaar aan te houden. Omdat ik destijds nog het idee had dat ik iets met muziek in mijn onderzoek wilde doen, deed je je best een link met autisme-onderzoek te vinden. Je bracht me in contact met de biologie, om eens te kijken of er een combinatie van zangvogel- en eye-trackingonderzoek mogelijk was. Dat dat er niet helemaal van is gekomen, zit me totaal niet dwars. De samenwerking van de afgelopen jaren heb ik als ontzettend prettig ervaren: ik heb altijd de vrijheid gevoeld om paden te bewandelen die ik interessant vond. Bijvoorbeeld bij de bezoeken die ik aan Lund bracht om eye-trackingmethodologie voor baby-onderzoek uit te diepen, of de grotere onderwijstaak die ik op me wilde nemen. Altijd wist ik dat jij achter me stond en erop vertrouwdde dat het goed kwam. Ook nu ik dit schrijf en mijn promotie in zicht is, doe je er alles aan om mij in je groep te houden. Dat waardeer ik enorm, en ik ben dan ook blij dat ik de komende jaren in Utrecht kan blijven! Chantal, van harte bedankt!

Ten tweede, Ignace, jij bent één van de grote redenen waarom ik deze baan heb aangenomen. Ik kende je al langer, gaf les in twee vakken die jij

coördineerde, en ik geloof dat we toen zelfs al aan het jammen zijn geweest. Hoewel je als mijn ‘dagelijks begeleider’ op een aantal documenten stond, heb ik het al snel niet als begeleiding, maar als samenwerking ervaren. Vanaf de eerste dag pushte je me elke keer de lat hoger te leggen, zowel in het onderzoek als in het onderwijs. Overdag hadden we kritische discussies, en ’s avonds dronken we bier, aten we burgers, en luisterden we goeie muziek. We zijn zelfs 2 maanden huisgenoot geweest in Noorwegen. Dan moet je wel met elkaar door één deur kunnen. De mooiste kans die je me gegeven hebt is om als tweedejaars AIO een college voor 400 studenten te geven in de cursus Experimenteren & Registreren 1. Dat was niet als kunstje om me af te laten gaan, maar omdat je het vertrouwen had dat ik dat aankon (ook al was ik bloednerveus). Inmiddels geven we al een tijd samen onderwijs en draaien we samen gave proeven. Ik kan me niet voorstellen dat dat heel snel gaat veranderen, en gelukkig maar!

Hoewel mijn naam op de kaft staat, is het hopelijk opgevallen dat er op de hoofdstukken nog een hoop extra namen staan. Zonder deze co-auteurs was dit proefschrift nooit gelukt. Carlijn en Tineke, dankzij jullie werk kon ik een vliegende start maken, waarvoor bedankt! Tim, zonder jou had dit proefschrift er echt heel anders uit gezien. Jij was verantwoordelijk voor het bouwen van de opstelling waarop ik onderzoek gedaan heb dat twee hoofdstukken in dit proefschrift opgeleverd heeft. Daarnaast heb ik met veel plezier de eye-trackertest met je opgezet en uitgevoerd! Zagen, schroeven, verven, programmeren, je zou je eigen handymanshow op TV moeten hebben. Ondanks dat je in Frankfurt zit, hebben we nog steeds een vruchtbare samenwerking. Bedankt! Diederick, de tweede keer dat ik in Lund was liet je terloops vallen dat je wel wilde helpen met de analyses voor het fixatiedetectie-algoritme dat ik had ontwikkeld, zodat we dat konden publiceren. Voor mij zat er op dat moment al een hele lange aanlooptijd in. Ik had allerlei technieken uitgeprobeerd, had ‘machine learning’-algoritmes geprogrammeerd die drie weken aan het analyseren waren en toch geen verbetering opleverden, en had uren aan baby-eye-trackingdata doorgespit. Ik was eigenlijk wel klaar met dat algoritme. Ik was blij genoeg dat het werkte. Ik was dan ook heel blij dat je aanbod te helpen. Het feit dat jij

## 12. Acknowledgements

nu als co-eerste auteur op dit stuk staat zegt dus wel wat over hoeveel werk je in dit project hebt gestoken, en hoe belangrijk je bijdrage was. Nadat het fixatiedetectie-algoritme gepubliceerd was gingen we meteen het volgende project aan, wat mij betreft bewijs genoeg dat het een goede samenwerking was! Gijs... Je thesis duurde zo 10 maanden in plaats van 5,5. En niet omdat je er de kantjes van af liep. Integendeel! Vanaf de eerste dag ben je met zo'n enthousiasme aan de slag gegaan, dat ik soms moeite had je bij te houden. Dat je als co-auteur op het laatste paper staat in dit proefschrift is meer dan verdiend: jouw kennis van de geschiedenis van de psychologie en je gave om de meest obscure filosofen aan te halen in onze discussies hebben er voor gezorgd dat ik tijdens onze samenwerking gigantisch veel geleerd heb. De gewoonte om over de interactieopstelling te praten, te filosoferen, bier te drinken en te jammen op één avond, houden we gelukkig nog steeds in ere!

Marcus, I remember dropping the idea of replicating your data quality study with infants somewhere in 2014. In a few short Skype sessions we put together a plan, and brought Richard in to discuss the analysis. In February 2015 I showed up in Lund with a large infant eye-tracking data set and we got to work. Richard, your lightning-fast lessons on linear mixed-effects models is the first reason for making this work in just one month! Marcus, you told me that I wrote faster than you could read, only to come back with a list of spot-on critique points in two hours. That's reason number two for making this work in such a short time! Richard, Marcus, tack så mycket!

Also a big thanks to all the proofreaders of the chapters in this dissertation: Ignace, Chantal, Nicolette, Tim, Marcus, Richard, Jeroen, Pap, Gijs, Lawrence, and Rudolf.

Ik heb de afgelopen jaren ook het genoegen gehad met een groep studenten samen te werken waarvan het werk niet direct in dit proefschrift terecht is gekomen (maar daarom niet minder cruciaal is geweest!). Lucas,

Daniel, Vivianne, Lawrence, Stefan (al was je eigenlijk de stagiair van Rudolf), Gijs, en Jordy, bedankt! Vivianne, je voerde een ‘creative challenge’ uit dat praktisch een heel bacheloronderzoek was in je eentje... Mijn dank is groot! Lawrence, ik kende je al van de cursus “*Leren Lesgeven in het Hoger Onderwijs*” en het was voor mij een groot plezier ook in je master weer samen te werken. Ik zie uit naar onze volgende filosofische discussie!

In het uitwerken van het onderzoeksvoorstel voor mijn AIO-periode heb ik met een aantal diverse groepen gediscussieerd. Johan, Sanne, en Thijs van de biologie, jullie waren erg enthousiast over mijn voorstel om sociale interactie bij zangvogels te onderzoeken. Ik heb hier veel van geleerd, ook al bleek het project uiteindelijk te ambitieus. Herbert en Yasin van methoden en statistiek, met jullie discussieerde ik over een analysemethode om een deel van de problemen beschreven in onze interactiestudies op te lossen. Ik zie uit naar het vervolg van onze samenwerking.

Nicolette, hoewel jij niet als co-auteur ergens in dit proefschrift staat, zou je het haast moeten zijn. Veel stukken heb je immers doorgelezen en bekritiseerd, en over nog meer stukken hebben we gediscussieerd. Je begon maar een paar maandjes later dan ik, en ik heb je dan ook altijd als AIO-maatje van het eerste uur beschouwd. Inmiddels ben je ook buiten het werk een hele goede vriendin van me geworden. Ik heb er vertrouwen in dat we elkaar ook over een paar jaar nog wel zien en spreken (ook al zouden we aan de andere kant van de wereld werken). Bedankt voor al je hulp de afgelopen jaren!

Ook alle andere directe collega’s ontzettend bedankt voor alle informele discussies door de jaren heen: Aliegriet, Maria, Rianne, Janna, Caroline, Renata, Bauke, Lilli, Malu, Jessica, Gonneke, Paula, Jacco, en alle meetassistenta’s in het KinderKennisCentrum. En zonder ondersteuning was ik helemaal nergens geweest: Emma, Cornelia, Jacobine, Maartje, Ria, Eveline, bedankt!

## 12. Acknowledgements

Collega's zijn er natuurlijk ook buiten onze onderzoeksgroep. Sterker nog, ik liep al een aantal jaren rond op de afdeling Psychologische Functieleer. Dat is allemaal te danken aan Chris P. en Stefan. Chris en Stefan leerde ik kennen toen ik in mijn tweede bachelorjaar het honoursprogramma in rolde. Daar is een groot deel van mijn interesse voor onderzoek ontstaan en ik mag dan ook van geluk spreken dat ik vanuit dit programma anderhalf jaar als onderzoeksassistent heb gewerkt voor Chris, Stefan, Tanja, en Maarten. Zowel mijn bachelorscriptie als mijn masterscriptie schreef ik vervolgens ook nog eens bij Chris, en samen met Chris en Stefan publiceerde ik voor het eerst. Stefan, inmiddels hebben wij weer samengewerkt aan een artikel, en Chris, wie weet gaan wij dat ook weer doen. Mijn mentoren van het eerste uur, bedankt!

Jonathan, mijn onderwijs- en klimmaatje. Ik baal ervan dat je inmiddels verhuisd bent naar de VU, maar dat mag de pret niet drukken. Niet alleen op werk kan ik met je lachen, je hebt me ook buiten het werk veel geholpen. Mijn dank is groot, maar niet te groot, want dan hou je er nooit meer over op. Vi sees snart! Rudolf, niet alleen hebben we prettig samengewerkt aan onze visuele zoekproef, de gesprekken waren ook altijd goed! Of het nu over tiny houses (ben je al aan het bouwen?) of de balans tussen werk en privé ging. Ik zie uit naar de volgende! Jeroen, met jou onderwijs geven gaat zo ontzettend soepel, gold dat ook maar voor het gamen op die arcadekast van je. Volgende keer versla ik baas nummer 3! Ik wil ook graag alle collega's in de 'perception'-meetings bedanken voor de input (voor zover nog niet genoemd en in de hoop dat ik niemand vergeet): Chris J., Susan (natuurlijk ook voor alle hulp bij het onderwijs), Remo, Surya, Jim, Sjoerd, Casper, Manje, Alessio, Serge, Barry, Ben, Wietske, Hinze, Marnix, Martijn, Jasper, Jelle, Paul, Joris. Tenslotte bedankt aan alle LLIHO'ers met wie ik elk jaar in blok 2 met gigantisch veel plezier les heb gegeven.

I consider myself lucky that my job has brought me to a number of different universities around the world. First of all, thanks to Kenneth for the warm welcome in Lund, the drive to Stockholm and Turku, and of

course the varmrökt lax! A number of the colleagues in Lund have already been thanked before, but I must not forget Alex and Raimondas for the wonderful times in Lund. Don't worry, I don't plan on staying away too long. Eelke and Morten, thanks for the ease with which Ignace and I could arrange our visits to the University of Tromsø, and Eelke in particular for the wonderful drive out to Sommarøy. Takk en bedankt! Alan, thanks for the opportunity of visiting the BAR lab at the University of British Columbia. Even though it was only a 2.5-week visit, I felt very welcome in the lab, and was glad to leave with a plethora of new ideas for our interaction setup. Tenslotte, Jenny bedankt voor het gemak waarmee ik opgeteld meer dan 2 maanden je huisgenoot was in Zweden en Noorwegen!

Een proefschrift schrijven gebeurt niet alleen op werk. Het is een klus waar ook buiten de universiteit een hoop tijd in gaat zitten. Gelukkig waren er genoeg mensen in de buurt om me vooruit te helpen of van afleiding te voorzien als dat nodig was. Marjoleine en Mosh, aan jullie de eer om bovenaan dit lijstje te staan. Jullie maken van Terwijde een huis waar ik me echt thuis voel. Bedankt voor alle aanmoediging, frutsel! Ik hoop dat ik die nog heel lang van je mag ontvangen! MJ, jij bent de persoon geweest op wie ik kon bouwen de afgelopen jaren! Bedankt voor alle hulp! Pap en Christine, Tim, Cindy et petit Erik, jullie zijn mijn thuis ver van huis. Elke keer als ik in Genève kom, heb ik het gevoel op vakantie te zijn. Bedankt voor alle rust die ik daar de afgelopen jaren heb kunnen vinden. Jelle, bedankt voor de hulp met het maken van de kaft en de gezelligheid de afgelopen jaren. Millad, Marloes, en Silvio, jullie bedankt voor al het leuks in goede tijden, maar nog veel meer voor alle hulp in minder goede tijden! Bedankt ook aan alle falers voor de vele mooie momenten de afgelopen 10 jaar. Ook al heb ik soms wat momenten gemist de afgelopen jaren, het is altijd thuishomen bij jullie. Martijn, Lawrence, en Kiki, hoewel jullie al verkapt in dit dankwoord staan, zijn jullie allang geen collega's meer, maar goede vrienden, dus bij deze: *\*maakt een willekeurige Freek-referentie\**.

Wie dit dankwoord van voor naar achter gelezen heeft zal misschien een tweetal namen missen. Sterker nog, ik denk dat als zij dit lezen ze zich

## *12. Acknowledgements*

afvragen waar ze nou in hemelsnaam blijven. Op de belangrijkste plek van het dankwoord. Mam, Sanne... Deze is voor de tijd op Denemarkenplein. Deze is voor jullie. Bedankt voor alles!

Roy Hessels, maart 2017

## **13. List of publications**

### 13. List of publications

#### **In this dissertation, published:**

- Hessels, R. S., Hooge, I. T. C., Snijders, T. M., & Kemner, C. (2014). Is there a limit to the superiority of individuals with ASD in visual search? *Journal of Autism and Developmental Disorders*, 44(2):443–451.
- Hessels, R. S., Cornelissen, T. H. W., Kemner, C., & Hooge, I. T. C. (2015). Qualitative tests of remote eyetracker recovery and performance during head rotation. *Behavior Research Methods*, 47(3):848–859.
- Hessels, R. S., Andersson, R., Hooge, I. T. C., Nyström, M., & Kemner, C. (2015). Consequences of eye color, positioning, and head movement for eye-tracking data quality in infant research. *Infancy*, 20(6):601–633.
- Hessels, R. S., Kemner, C., van den Boomen, C., & Hooge, I. T. C. (2016). The area-of-interest problem in eyetracking research: A noise-robust solution for face and sparse stimuli. *Behavior Research Methods*, 48(4):1694–1712.
- Hessels, R. S., Hooge, I. T. C., & Kemner, C. (2016). An in-depth look at saccadic search in infancy. *Journal of Vision*, 16(8):10.
- Hessels, R. S., Niehorster, D. C., Kemner, C., & Hooge, I. T. C. (2016). Noise-robust fixation detection in eye movement data: Identification by two-means clustering (I2MC). *Behavior Research Methods*.
- Hessels, R. S., Cornelissen, T. H. W., Hooge, I. T. C., & Kemner, C. (2017). Gaze behavior to faces during dyadic interaction. *Canadian Journal of Experimental Psychology*.

#### **In this dissertation, submitted:**

- Hessels, R. S., Holleman, G. A., Cornelissen, T. H. W., Hooge, I. T. C., & Kemner, C. (2017). Eye contact takes two – capturing gaze behavior of subclinical autism and social anxiety in dyadic interaction. *Submitted*.

#### **Not in this dissertation:**

- Paffen, C. L. E., Hessels, R. S., & Van der Stigchel, S. (2012). Interocular conflict attracts attention. *Attention, Perception & Psychophysics*, 74(2):251–256.
- Niehorster, D. C., Cornelissen, T. H. W., Holmqvist, K., Hooge, I. T. C., & Hessels, R. S. (2017). What to expect from your remote eye tracker when participants are unrestrained. *Behavior Research Methods*.

Cousijn, J., Hessels, R. S., Van der Stigchel, S., & Kemner, C. (2017). Evaluation of the psychometric properties of the gap-overlap task in 10-month-old infants. *Infancy*.



## **14. Curriculum vitae**

## 14. Curriculum vitae

Roy Hessels was born on April 12th 1990 in Voorburg, the Netherlands. After graduating from a bilingual program at high school (“tweetalig VWO”) in 2007, he started his Psychology Bachelor at Utrecht University. During his Bachelor, he worked as a research assistant for 1.5 years conducting psychophysics research. In 2010 he graduated with honours, and began his Master in Applied Cognitive Psychology at Utrecht University. During his Master, Roy conducted research on rhythm perception at the School of the Arts Utrecht (HKU), and investigated subjective rhythm for the purposes of Brain-Computer Interfacing using EEG at the Radboud University Nijmegen. In 2011, he graduated cum laude from his Master’s program.



Following his graduation, Roy worked as an applied psychologist at the School of the Arts Utrecht for 7 months before starting as a research assistant at Utrecht University in the group of Chantal Kemner. During the first year, he wrote a proposal for a three-year PhD-trajectory, which was approved to start in January 2014. During his PhD, Roy spent time at Lund University, the University of Tromsø, and the University of British Columbia. Next to his research, he was involved as a teacher in setting up a number of eye-tracking courses, and was responsible for introducing *Bring-Your-Own-Device* into the Psychology Bachelor program. In 2015 and 2016, Roy was nominated by the students for the junior Psychology teacher of the year award. In 2016, he was awarded the Maarten van Son junior teaching award. Roy is currently employed as a post-doc at Experimental Psychology and Developmental Psychology at Utrecht University.