Janina Saarnio

EXAMINING THE ROLE OF JOINT VISUAL ATTENTION AND TACTILE ARTEFACTS IN ENHANCING LONG-TERM MEMORY

JYVÄSKYLÄN YLIOPISTO
INFORMAATIOTEKNOLOGIAN TIEDEKUNTA
2024

TIIVISTELMÄ

Saarnio, Janina
Examining the role of joint visual attention and tactile artefacts in enhancing long-term memory
Jyväskylä: Jyväskylän yliopisto, 2024, 76 s.
Kognitiotiede, pro gradu -tutkielma
Ohjaaja(t): Kujala, Tuomo; Simola, Jaana

Nykyinen tietotyö sisältää useita kokouksia, joiden käytäntöjä ei ole optimoitu. Pääasiassa suulliseen ja tekstuaaliseen tiedonvaihtoon perustuvat kokousmenetelmät ovat alttiita väärinkäsityksille monimutkaisia syy-seuraussuhteita ratkottaessa. Tässä tutkielmassa tutkittiin taktiilisten artefaktien käyttöä visuaalisina kommunikaation apuvälineinä osana strategian laadintaa. Visuaaliset kommunikaatiomenetelmät tarjoavat näkyviä viittauspisteitä yksilön ajatusmalleihin, jotka edesauttavat jaetun yhteisymmärryksen luomisessa. Kommunikoinnin ja strategian laadinnan tehostamisen lisäksi artefaktien oletettiin kiinnittävän 3–4 hengen ryhmissä toimivien osallistujien (jaetun) visuaalisen tarkkaavaisuuden, jota tutkittiin silmänliikettä seuraavien lasien avulla. Jaetun visuaalisen tarkkaavaisuuden (JVT) on tutkittu parantavan yhteistyön laatua ja edistävän oppimista. Tässä tutkielmassa tutkittiin (jaetun) visuaalisen tarkkaavaisuuden vaikutusta osallistujien narratiiviseen ja visuaaliseen pitkäkestoiseen muistiin yhteistyössä laaditusta strategiasta, jota tutkittiin puolistrukturoidun verkkohaastattelun avulla. Narratiivisen pitkäkestoisen muistin määriteltiin koostuvan artefaktien avulla luodusta suullisesti selitettävissä olevasta strategiasta, kun taas visuaalinen pitkäkestoinen muisti koski yksittäisiä artefakteja ja niiden merkityksiä. Lisäksi silmänliikedatasta analysoitiin, kuka osallistujista aloitti kunkin jaetun visuaalisen tarkkaavaisuustapahtuman, jonka on tutkittu vaikuttavan muun muassa kuvien tunnistusmuistiin. JVT:n ja sen aloittamisen vaikutusta pitkäkestoiseen muistiin analysoitiin binäärisellä logistisella monitasomallilla, jossa havaintoyksikkönä käytettiin yksittäistä visiittiä eli yhtäjaksoista katsetta tarkasteltavalla alueella. Yksinomaa visuaalisen tarkkaavaisuuden vaikutusta pitkäkestoiseen muistiin analysoitiin lineaarisella monitasomallilla, jossa jatkuvana riippuvana muuttujana oli yksittäisen visiitin kesto. Tutkimuksen mukaan JVT ei paranna narratiivista tai visuaalista pitkäkestoista muistia. Sen sijaan JVT:n aloittaminen parantaa visuaalista pitkäkestoista muistia. Tulosten mukaan yksittäisten fiksaatioiden lukumäärä visiitin aikana lisää JVT:n ja sen aloittamisen todennäköisyyttä, kun taas pidempi visiitin kesto heikentää näiden todennäköisyyttä. Tutkimustulokset korostavat dynaamisen ja aktiivisen visuaalisen tarkkavaisuuden merkitystä tiimityöskentelyssä pitkäaikaisen muistin parantamiseksi.

Asiasanat: jaettu visuaalinen tarkkaavaisuus, tarkkaavaisuus, pitkäkestoinen muisti, taktiiliset artefaktit, katseenseuranta

ABSTRACT

Saarnio, Janina
Examining the role of joint visual attention and tactile artefacts in enhancing long-term memory
Jyväskylä: University of Jyväskylä, 2024, 76 pp.
Cognitive Science, Master's Thesis
Supervisor(s): Kujala, Tuomo; Simola, Jaana

Daily routines in knowledge-driven work are built around meetings whose practices are not optimized. Relying mainly on verbal and textual exchange of information, misunderstandings can easily occur when the aim is to solve complex cause-and-effect relationships. This thesis study leveraged tactile artefacts as visual communication aids in strategy development. Visual communication methods are said to provide visual points of reference for individual mental models, which facilitate the building of common ground and shared understanding. Besides enhancing communication and strategy development, tactile artefacts were assumed to attract (joint) visual attention of the participants working in groups of 3-4 people, which was studied using mobile eye-tracking. Previous research has found that joint visual attention (JVA) enhances collaboration quality and learning. This thesis examined the effect of (joint) visual attention on visual and narrative long-term memory of the collaborative developed strategy, which was assessed with a half-structured online interview. The narrative long-term memory referred to the verbally explainable strategy, which was iteratively developed with the help of tactile artefacts, while the visual long-term memory pertained to individual artefacts and their assigned meanings. Additionally, the initiators of JVA events were examined from the eye-tracking data, as initating JVA has been shown to enhance recognition memory. The effect of JVA and initiating JVA on long-term memory was analysed with the binary logistic mixed-effects model, using gaze visit—defined as detected continuous gaze on the area-of-interest—as the unit of analysis. The sole effect of visual attention on long-term memory was analysed with a linear mixed-effects model with visit duration as a dependent variable. The results indicate that JVA does not enhance narrative or visual long-term memory. However, initiating JVA does enhance visual long-term memory. The eye-tracking metrics also revealed that the number of fixations during a visit increase the odds of both JVA and initiating JVA, whereas longer visit duration decrease the odds of both. The findings of this study highlight the importance of dynamic and proactive visual engagement and participation in enhancing long-term memory in collaborative settings.

Keywords: joint visual attention, visual attention, tactile artefacts, long-term memory, mobile eye-tracking

FIGURES

TABLES

CONTENT

1 INTRODUCTION

It has been reported that on average executives spend at least 10 hours per week in meetings; however, only 10 percent of them have not received training on anything more than how to build a meeting agenda (Napier & Sharp, 2019). Ineffective meetings have direct costs in terms of salary and time wasted. In addition, they also cause indirect costs associated with employees' stress, frustration and fatigue from the unbeneficial meetings and the work allocated for preparing them (Allen et al., 2022).

While the use of facilitators in meetings has become common to improve collaboration quality, the facilitation methods are often based on verbal or textual exchange of information. Although structured group discussions have been shown to promote proactive communication and inhibit ineffective meeting behaviours compared to unmanaged (free) social interactions (Lehmann-Willenbrock et al., 2013), the risk of miscommunication is always present (Paxton et al., 2021). For example, inaccurate estimations of the listener's needs and previous knowledge, along with the tendency to reduce one's own cognitive effort by providing less information, can increase communication ambiguity and easily lead to misunderstandings (Paxton et al., 2021).

Successful social interaction through verbal communication requires adapting to each other's linguistic behaviours to achieve shared understanding (Fusaroli et al., 2012, Garrod & Pickering 2004). This process is commonly referred to as building and maintaining an implicit common ground, which has so-called associated grounding costs depending on the medium of communication and its affordances—which refer to the perceived opportunities for interaction that a certain communication medium provides (Homaeian et al., 2021). For example, in face-to-face conversation the grounding costs are typically lower compared to phone calls because also non-verbal cues, such as facial expressions, gaze and gestures, can be used.

In comparison to textual and verbal communication, visual communication offers distinct advantages. For example, it can help to overcome language barriers, capture attention, and convey information more clearly. Typical modes for visual communication in business strategy and development work include infographics, charts, graphs, and presentations. Research indicates that these visual tools enhance information retention and understanding among diverse audiences, facilitating more effective communication of complex concepts and data (Lankow et al., 2012). The effectiveness of leveraging visual communication in collaboration can be attributed to the principles outlined in the dual coding theory. This theory, proposed by cognitive psychologist Allan Paivio (2013), suggests that humans process information in two distinct but interconnected ways: verbally and visually.

Expanding beyond traditional visual mediums, *tactile artefacts* as movable physical objects, represent an alternative medium for visual communication. Besides facilitating our own cognitive processes, such as learning and sense-making, tactile artefacts have been used to communicate our knowledge for others (Zenk et al., 2021, Teh et al. 2021). This can be particularly useful in collaborative settings, where tactile artifacts can help make abstract ideas more concrete and clarify concepts that might be difficult to express with words alone. To date, LEGO® SERIOUS PLAY® (LSP) bricks are the most famous tactile artefacts used to support collaborative work in the business world. According to the Serious Play Community (n.d.), the method is based on iterative building and ideation to build a common understanding to the problem at hand, where each participant constructs "a metaphorical model to tell their own story" with the help of LSP bricks. Several studies have reported the value of these tactile artefacts in teamwork. For example, the use of LSP bricks in innovative workshops have been rated as more enjoyable, active and inspiring than traditional moderation cards (Zenk et al., 2021). Other empirical studies have concluded that using LSP bricks can have a positive role for developing psychological safety (Wheeler & Passmore, 2020) and creating mutual understanding by stimulating intuitive playful modelling (Schulz et al., 2015).

In addition to tactile artefacts enabling participants to construct physical representations of their ideas and concepts (Roos & Victor, 2018), they can presumably direct shared attention among collaborators. In fact, *joint visual attention* (JVA), defined as shared overt visual attention to an area-of-interest (AOI), has been shown to support learning (Schneider & Pea, 2013) and increase the quality of product (Fındık-Coşkunçay & Çakır, 2022), interaction (Jermann et al., 2011) and collaboration (Schneider et al., 2018, Schneider & Bryant, 2024). Enhancing joint visual attention could also address the issue of ineffective meeting practices because research indicates that JVA improves both visual working memory (Gregory & Jackson, 2019, Gregory & Jackson, 2017, Gregory & Kessler, 2022) and to some extent, long-term memory (Kim & Mundy, 2012, Dodd et al. 2012). Consequently, tactile artefacts emerge as a promising solution enhancing collaboration and participants' recollection of meeting outcomes to facilitate effective strategy development and execution.

The purpose of this master thesis is to study if and how tactile artefacts used in face-to-face interaction enhance joint visual attention and individual long-term memory of the group's outcome. Although it is quite well established that the reflexive act of gaze following, which refers to the natural tendency of humans to automatically shift their gaze in the direction where another person is looking at, enhances memory (Gregory & Jackson, 2019, Großekathöfe et al., 2020, Yin, 2022), the effect of joint visual attention to individual long-term memory (LTM) is yet little studied in real-life social interactions. Additionally, gaining a deeper understanding of the effects of visual communication aids on joint visual attention, and their subsequent impact on long-term memory, could benefit various fields, such as education and software development.

This thesis was conducted as part of a Business Finland-funded Research to Business project at Aalto University (Department of Industrial Engineering and Management), with no commercial interests influencing its production.

1.1 Research questions

The main research question is "If and how does (joint) visual attention to tactile artefacts in face-to-face collaboration enhance individual long-term memory of the group's outcome?". Although joint visual attention is the main interest of this study, also individual attention to tactile artefacts is researched.

Other research questions are related to different types of memory because the developed strategy as the group's outcome can be understood in two different, but related ways. The participants can either have a strong visual long-term memory of the artefacts used during the collaborative workshop, or a strong narrative long-term memory related to the jointly built strategy with the help of artefacts. Therefore, the second research question is "If and how does (joint) visual attention enhance visual and/or narrative long-term memory of group's outcome?". It is hypothesised that joint visual attention enhances both visual and narrative long-term memory of the group's outcome because the used artefacts are thought to 'embody' the details of the jointly built strategy.

1.2 Structure

The second chapter of the thesis entails a literature review on the topic. The third chapter covers the hypothesis, experimental design, and analysis methods, including the rationale behind operationalizing the variables. The fourth chapter presents the results of the study, while the fifth chapter discusses the obtained results. The final chapter concludes the study.

2 HUMAN ATTENTIONAL MECHANISMS IN SOCIAL INTERACTION

2.1 Attention and memory

Attention refers to the selection processes of our cognitive system that prioritises certain stimuli for further information processing, which is a necessary mechanism due to our limited cognitive capacities and resources (Styles, 2005, Carrasco, 2011). A consensus exists that attention cannot be regarded as a unitary process because it describes a diverse set of behavioural and cognitive phenomena (Cohen, 1993, Carrasco, 2011, p. 1517). In fact, it has been argued that attention should be studied as a set of attentional processes that metaphorically acts as filters or gates for further information processing in the brain. This metaphorical description of attention generally refers to the concept of *selective attention* (Cohen, 1993, p. 3).

The attentional processes encompass attention, perception, and memory, which are often perceived as distinct yet interconnected cognitive functions or stages of cognition (Styles, 2005). Perception has been associated with both bottom-up sensory processing, where many of its initial stages occur unconsciously and automatically (Styles, 2005, p. 7), as well as with the top-down endogenous control of attention (Cohen, 1993, p. 26). Perception and attention are intertwined, as while we continually perceive our surroundings, we cannot actively hear a distant conversation without directing attention to it. In essence, attention serves as both 'an active agent' and an outcome of sensory processing. For example, it has been expressed that we ultimately perceive the object of attention by binding its visual features together (Styles, 2005, p. 6). Consequently, visual attention has been described as the process that transforms looking into meaningful seeing (Carrasco, 2011, p. 1484).

Instead of a sole information store, memory can also be defined as a dynamic attentional process that entails encoding and retrieving information (Styles, 2005), and sometimes forgetting it. In the realm of attentional processes, memory plays a crucial role in shaping and directing attention. For example, the cognitive system relies on memory to store and retrieve information about current and past experiences, which in turn is known to influence the allocation of attention (Hirschstein & Aly, 2023).

The different memory processes are usually defined by their duration. First, the very brief sensory memories, such as an iconic memory for visual information, act as a buffer for selective attention (Styles, 2005). The concept of short-term memory (STM) varies in different theoretical models binding attention and memory, some regarding it as a separate and passive store, whereas others see it more closely related to working memory (WM) (Norris, 2017). A consensus exists that the WM holds the information which is arriving from perceptual processing or that is retrieved from long-term memory (LTM) stores. Although both short-term memory and working memory have a limited capacity, the information is said to stay active in working memory only as long as it is consciously attended, indicating a close relationship between the concepts of working memory and attention (Styles, 2005, Norris, 2017). For example, a study on individual differences in top-down controlled attention and working memory have shown a correlation between working memory capacity and the efficiency of controlled attention (Awh et al., 2006).

Memory is also regarded as the by-product of attention as the more we direct our attention to certain stimuli, the more likely we are to remember them (Styles, 2005, p. 9, Cohen, 1993, p. 392). Once the attended stimuli are selected, attention is closely involved in the information processing, which is referred to as *sustained attention* (Styles, 2005, p. 6). Sustained attention is needed for information consolidation into our long-term memory (deBettencourt et al., 2021). The long-term memory entails memory for all stored knowledge and it theoretically has an unlimited capacity. There are different types of LTM; the declarative, explicit long-term memory includes semantic memory for facts; episodic LTM stores personal experiences, whereas the procedural, implicit LTM stores actions (Styles, 2005, p. 9).

Collaborating with tactile artifacts in a narrative context—where the jointly built strategy is verbally explained following storytelling principles—requires considering also other dimensions and aspects of memory that might support long-term memory consolidation. For example, according to *schema theory*, a schema as a mental framework represents general knowledge about a particular domain and determines what information will be encoded within that domain (Alba & Hasher, 1983). In the context of narratives, schema theory suggests that schemas play a significant role not only in interpreting the events of a narrative based on existing knowledge, but also in encoding, storing, and retrieving those events—as well as the event details (Masís-Obando et al., 2022). The possible effects of tactile artefacts and narratives on memory will be discussed more in detail in section 2.4.2.

2.2 Selective attention and its related concepts

The environment is full of information which cannot be processed all at once. To make sense of our surroundings and navigate through this sea of data, we need to rely on selective attention. To understand what has been guiding the 21st-century (neuro)psychological visual attention research, it is beneficial to understand the historical context and development of the selective attention theory.

2.2.1 Short history of selective attention theories

In the 1950s experimental psychologists started to study different filtering mechanisms related to selective attention and attention allocation. According to Ronald Cohen, an author of *The Neuropsychology of Attention* (1993), the theory of selective attention is originally derived from information theory, which suggests that the nervous system can select certain information channels while filtering something out (Cohen, 1993, p. 23). However, the initial assumption of an automatic and early all-or-none filtering process of selective attention led the researchers in the late 1950s to review the bottleneck theory, as automatic filtering of sensory inputs before pattern recognition could not explain certain observed phenomena, such as why individuals tend to hear their own name over others (Cohen, 1993).

Initially, Donal Broadbent (1958) reviewed his early filtering theory by suggesting a two-process model of attentional selection, where 'pigeonholing' would complement the filtering process by sorting out stimuli into response categories (Cohen, 1993, p. 24-26). The active and conscious act of perception as an attentional process was then introduced by Ulric Neisser (1967), who aimed to end the debate about the location of the attentional filter (Cohen, 1993, p. 51). Neisser adopted Frederik Bartlett's (1932) term 'schemata' to describe that perceivers do not simply filter information but actively gather data in an alignment with their present expectations, intentions and past encounters, referring to the top-down control of attention (Cohen, 1993, p. 26). Neisser understood cognition as constructive, and in relation to memory, he regarded those schemata as cognitive structures also "controls the fate of stored information" (Neisser, 1967, p. 287).

In the 1970s, the research on the selectivity of attention shifted to primarily focus on the automatic vs. controlled processing of attention (Cohen, 1993, p. 35), also referred to as bottom-up vs. top-down control of attention. The experimental studies on automatic control led researchers to propose a new theory on human information processing, emphasising the role of memory in attentional processes (Cohen, 1993, p. 39). According to Cohen (1993, p. 39) this structuralist point of view regarded memory as a collection of associated nodes that contain many information elements, which all become activated once one element gets activated. Daryl J. Schneider and Walter Shiffrin (1977) introduced two different types of memory stores to define the state of these nodes; the long-term store consisting of inactivated nodes, and the limited short-term store consisting of activated nodes (Cohen, 1993, p. 39). In addition, they (1977) also defined a concept of control processes to describe different conscious attentional processes that will activate a series of nodes to accomplish a task. In fact, another role of the short-term store was to provide "a workspace for control processes" in addition to storing relevant information coming from sensory inputs (Cohen 1993, p. 39).

Highlighting the role of memory in attentional processes naturally led into discussions about the capacity and resources of our attentional system. In other words, besides the automatic bottom-up and volitional top-down attentional control, the capacity or the total number of processing resources of the nervous system seemed to play an important role in governing attentional allocation (Cohen 1993, p. 41). Daniel Kahneman (1973, p. 8) developed the capacity theory of attention, which suggests that there are general limits to performing mental work. According to Kahneman, attention is allocated according to the available resources and the demands of the task at hand, stating that attention always requires some amount of cognitive effort (Kahneman, 1973). Except for the preattentive process of sensory filtering, Kahneman (1973, p. 122) rejected the filter theory of attention. Following Anne Treisman's famous work on visual attention, he argued for a parallel information processing and divided attention. Inherited from Kahneman, nowadays attention is often characterised as pools of processing resources, where attentional resources are either divided or directed all into one task (Styles 2005, p. 6).

In terms of the early filtering theory of selective attention, it has become apparent that selective attention plays a role both during early sensory processing and later post perceptual stages (Awh et al., 2006). It is also widely accepted that attention can be allocated either voluntarily through top-down control, and automatically through bottom-up control. However, understanding the intricate interplay between these mechanisms is not straightforward. In the next section, the selectivity of visual attention will be discussed more in detail, focusing on how the allocation of visual attention can be influenced by complex factors associated with environmental cues and features, which forms an important part of this thesis.

2.2.2 Selectivity of visual attention

Regarding visual attention allocation and 'selective looking' it is known that certain features, such as luminance and colour, effectively capture attention compared to less salient stimuli. The feature-based attention theory aims to explain the selective mechanisms by which attention is directed to specific features of objects (Theeuwes, 2013). The feature-based attention theory has emerged from visual attention studies, which often involve visual search tasks. The feature-based neuronal tuning of visual attention is thought to be likely governed by top-down volitional control because it has been widely observed that when participants are informed about the colour of the target object, their search prioritises that specific feature (Theeuwes, 2013). However, the effectiveness of luminance and colour in guiding attention is known to vary depending on the task, with their impact attributed to both bottom-up sensory processes and top-down cognitive factors (Theeuwes, 2013).

Besides allocating attention based on the visual features of objects, visual attention can also be directed to a specific location of the environment. The Posner's cueing paradigm (1980) is a classic experimental paradigm that demonstrates the role of spatial attention in guiding visual processing. In the experiment, the task of the participants is to look at a visual display containing a central fixation point (Frischen et al., 2007). Prior to presenting the target at the periphery, a cue is presented either in (or towards) a valid or invalid location. The results show that the participants are faster at detecting targets at the validly cued locations, suggesting that attention can be directed to the cued location (Theeuwes, 2013). This finding is referred to as space-based guided theory, or spotlight theory of attention, which posits that attention operates like a spotlight, selectively illuminating specific areas of the visual field (Styles, 2005, p. 75).

The feature-based and the space-based visual attention theories were expanded with object-based selective attention theory by John Duncan in 1984 (Duncan, 1984, Chen, 2012). Duncan followed Neisser's (1967) theory of focal, serially distributed attention to objects. He discovered that participants were able to select one of the two items from the same location, and two judgments concerning the same object could be made without the loss of accuracy (Duncan, 1984). The object-based selective attention challenges the feature-based attention theory by arguing that once the attention is directed to a certain feature of an object, the whole object will get selected—not only its certain feature (Cavanagh et al., 2023).

The selective attention theory is crucial for understanding how humans process information in their environment. Nowadays there is a consensus that visual attention can be directed either towards specific locations in the scene or directly to preattentively segmented objects (Soto & Blanco, 2004). The space-based, feature-based and object-based attention allocation processes are recognized to interact and influence each other during attentional tasks (Soto & Blanco, 2004). More recently, it has been also proposed these allocation processes are not entirely independent (Cavanagh et al., 2023). In addition, the famous Posner's cueing paradigm has shown that attention can be allocated overtly by directing the gaze towards the attended location or stimuli in the environment, or covertly by directing attention external stimuli without any eye-movements (Carrasco, 2011, Styles, 2005, p. 76). The next section focuses on visual perception and information-processing mechanisms that are predominantly driven by continuous eye-movements.

2.2.3 Human visual system and eye movements

The light enters the human eye through the cornea and lens. The neural component of the eye responsible for receiving stimuli is called the retina, which is located at the back of the eye (Styles, 2005, p. 49). The retina contains two types of photoreceptors; rods and cones, that are specialised to receive different ranges of light intensity. The rods are more numerous and sensitive to light than cones. Cones are responsible for conveying colours, and they are densely packed in fovea, which is often called "the centre of the eye's sharpest vision" (Frintrop, 2006).

These anatomical constraints of the human eye explain the reason for various types of eye movements; specifically, the fovea must be directed toward the stimulus for a clear vision (Rayner, 2009).

The four main types of eye-movements that guide visual search and attention allocation are saccades, stabilising movements, smooth pursuit movements and vergence movements (Rayner, 2009). Saccades are rapid, ballistic eye-movements whose function is to redirect gaze and scan specific features in the environment (Styles, 2005). Although the vision is suppressed during saccades, the cognitive processing continues. It is believed that attention precedes a saccade (Rayner, 2009) as the 'ballisticity' of saccades means that their direction cannot be corrected once they are initiated. The amplitude of the saccades (measured in degrees or minutes of arc) varies according to the task, higher amplitude referring to greater distance, which also correlates with the duration of the saccade (Rayner, 2009).

Although saccades are needed to maintain a coherent picture of the environment, visual information is primarily acquired during the slow stabilising eye-movements, generally referred to as *fixations* (Styles, 2005, p. 74). Fixations allow the visual system to gather detailed information about the objects or scenes being observed. Fixations keep the gaze stationary with two reflexes: the vestibulo-ocular reflex (VOR) compensating the head movements and the optokinetic reflex (OKN) stabilising a moving image on the retina. According to Holmqvist et al. (2011, p. 22) the name "stabilising eye-movement" refers to the fact that even during fixations the eyes are not completely still, but they perform three types of micro-movements: tremors, microsaccades and drifts. Tremors are involuntary small and rapid vibrations around 90 Hz that are governed by α-motoneurons of the brain stem nuclei. Their role is unclear (Holmqvist et al., 2011, p. 22), but it has been studied and suggested that the amplitude and frequency of ocular tremor movements predict fatigue (Lyapunov et al., 2022). Microsaccades are similar to regular saccades, except they are smaller and occur during fixations, which gives them a name of "fixational saccades" (Krekelberg, 2011). Finally, visual acuity is dependent on ocular drifts that are slow and irregular eye-movements (Clark et al., 2022).

Fixations are typically measured and reported by their average duration (ms), number (N), proportion (%) and rate or frequency in a certain AOI (Holmqvist et al., 2011, p. 399). The fixation rate or frequency, calculated by dividing the number of fixations by the trial period, indirectly includes saccade and blink durations (Holmqvist et al., 2011, p. 416). In addition to reporting the proportion of fixations in a certain AOI, the fixations can also be categorised by their duration and reported as proportions. Based on the argument that fixation durations are an indicator of different type of mental processing, Holmqvist et al., (2011, p. 416) give an example of segmentation, where fixation proportions in between 150 ms and 900 ms could be associated with cognitive processing, while the remaining fixations could be categorised as "express or overlong staring".

Rayner (2009) also highlights that although the neural circuitry that controls our eye-movements is the same across tasks, the main eye-movement patterns (fixation durations and saccade lengths) vary. For example, fixation durations are generally longer in scene perception (260-330 ms) than in silent reading (225-250 ms). During scene perception, viewers are also known to fixate only the most informative parts of the scenes that are often studied as scan paths. During the first fixation to the scene already a vast amount of information is acquired that is believed to guide the consequent fixations (Rayner, 2009). In addition, besides the average fixation duration, the number of fixations matter, as every fixation to an object of interest has an effect on recognition memory (Tatler et al., 2005).

The early eye-mind hypothesis (EMH) states that longer average fixation duration equals deeper processing. According to Holmqvist et al. (2011, p. 382) this hypothesis has been proven in reading tasks, where complicated texts or infrequent words cause longer fixations, in general, during more demanding tasks. However, they also point out that the EMH is task dependent and does not always have a one-to-one-correspondence. For example, longer fixations can also be an indicator of daydreaming (p. 382).

Eye-movement patterns also vary in face-to-face interactions compared to individual visual search or reading tasks (Amati & Brennan, 2018). In fact, many studies on gaze behaviour during social interaction have focused more on the dynamics of gaze behaviour and to the regulatory functions of the eye gaze i.e., in relation to speech turns instead of only measuring the individual fixation durations and saccade amplitudes per se (Brône & Oben, 2018). The next chapter focuses on gaze behaviour in social interactions by introducing the gaze cueing and following phenomena that have been studied to effectively guide our visual attention—and eye-movements.

2.3 Joint visual attention

As stated in the beginning, long-term memory can be understood as the by-product of attention and therefore attention and memory are often understood as intertwined. However, as deBettencourt et al. (2021) highlight, in long-term memory research attention is often treated as a “monolithic cognitive construct” without considering its subcomponents and their absence, which could better explain memory failures. For example, memory failures could be related either to misdirected spatial attention, which can be at least partially measured with eye-tracking technology, or due to the mind wandering effect caused by failed sustained attention, which is more difficult to detect (deBettencourt et al., 2021).

Long-term memory has famously been described as having an unlimited capacity with high fidelity (Brady et al., 2008), and it has been found that top-down guided attention enhances visual long-term memory (VLTM) (Sasin & Fougnie, 2021). By analysing the effect of related-context non-targets and salient distractors in a visual memory task, Sasin and Fougnie (2021) found that VLTM is not only dependent on the amount of attention because the related-context non-targets were better remembered than the salient distractions, which presumably drew attention through bottom-up control.

An intriguing phenomenon related to memory enhancement through the bottom-up control of attention is a gaze cueing effect (Dodd et al., 2012). Gaze cues are social cues that have been studied to effectively guide observers' visual attention, a phenomenon named as gaze following (Yin 2022, McKay et al., 2021). Numerous studies have found that gaze cues override the top-down attention control, which means that social signals are difficult to ignore (Großekathöfer et al., 2020, Cole et al., 2015). For example, a neurological study shows that compared to non-social cues (such as arrows), gaze cues do not elicit event-related potential (ERP) components related to voluntary control of attention, called the early direction of attention negativity (EDAN) (Nummenmaa & Calder, 2009). Another study by Großekathöfer et al. (2020) discovered that even when participants are asked to scan and remember the whole naturalistic environment, gaze cues effectively guide the attentional exploration.

Responding to gaze cues as an act of gaze following and engaging in social attention is regarded as a socio-cognitive, evolutionary, process that provides people important information about their surroundings and contributes to higher order social interaction skills (Sun et al., 2017, McKay et al., 2021). Eye-tracking research has revealed that eyes are the most (and first) attended facial feature (Thompson et al., 2019). As Hessels et al. (2023) state, eyes have a dual role of both information gathering and information signalling to others. Sensitivity to eye-gaze is a key for the development of social cognition from early life (Frischen et al., 2007). Even five-month-old infants can already detect very small horizontal deviations (5°) of eye gaze, although the perception of gaze direction and engaging in joint attention does not occur before the age of three (Frischen et al., 2007). In (neuro)psychological research joint visual attention refers to the common point of reference that develops before social cognition and language (Mundy & Newell, 2007). The effect of gaze following has been shown to decline by age (Slessor et al., 2008), but not without controversy (Kuhn et al., 2015). Additionally, another research has found that females have an advantage in gaze following compared to males, which has been explained by the "extreme male brain" hypothesis related to autism (Bayliss et al., 2005). In general, females have been suggested to have an advantage in judging nonverbal cues (Hall, 1978), which can be regarded as a prerequisite for joint visual attention.

In fact, besides perceiving the gaze direction of others by detecting the high contrast orientation of the pupil and iris relative to the large regions of white sclera, also higher-level factors, such as the assumed gaze direction towards an object instead of an empty space (Frischen et al., 2007), and the interpretation of the successive action of the observed actor (Perez-Osorio et al., 2015), guides gaze following. A key intention to follow another's gaze direction requires understanding that humans often direct their visual attention to their actions. For example, prior to developing the ability for engaging in joint visual attention, infants often discern the intentions of their caregivers by looking from their caregivers' head to their hands (Frischen et al., 2007). It has been also discovered that gaze following only occurs when the gazers are known or believed to be intentional, and therefore, gaze following is believed to be part of a larger human intentionality detection system (Wiese et al., 2012).

The act of gaze following as an overt orientation of visual attention according to the gaze cue is commonly studied with modified Posner's spatial cueing task. In this paradigm, a gaze cue is applied as a predictive or non-predictive cue which is presented equally often either towards the target or away from the target before the target stimulus is presented (see example in Figure 1) (McKay et al., 2021, Yin, 2022). In some versions of the experiment, a neutral directional precue is presented gazing towards the participant before the directional gaze cue is presented, although no direct effects have been reported by applying the direct gaze (McKay et al., 2021).

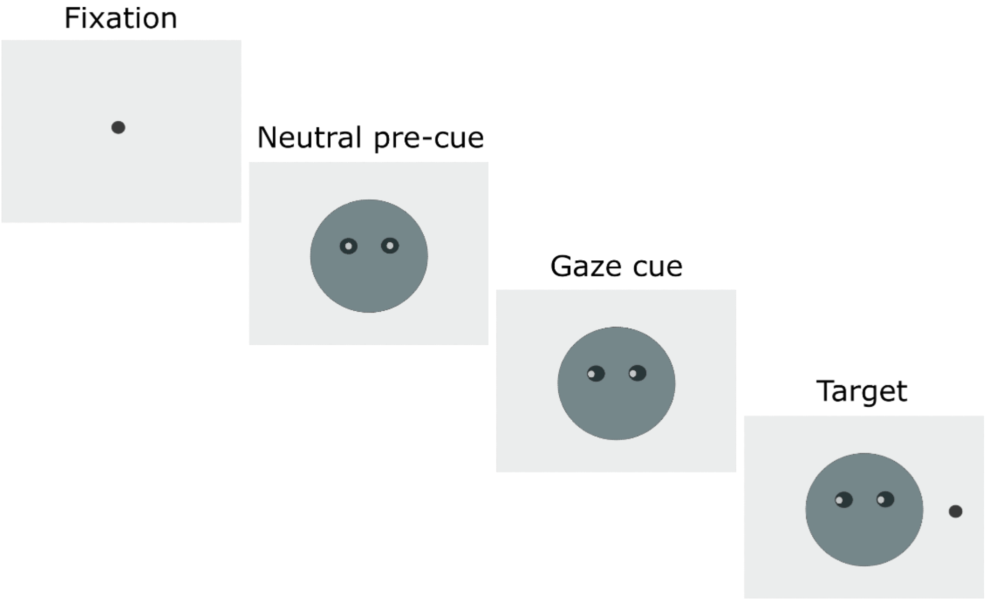


Figure 1. Traditional gaze cueing paradigm.

However, the traditional gaze cueing paradigm has been criticised for its lack of ecological validity, as the real-world visual environment is full of competing information (Großekathöfer et al., 2020). Therefore, the results from these classical gaze cueing paradigms are not directly applicable to the social cognition research, as the paradigm rarely entails real, intentional social interaction (Caruana et al., 2017). The first study on gaze-following in real-life on human crowds was conducted by Milgram et al. (1969) who set a stimulus group standing and staring up into a building window in a crowded New York city street. They observed and measured the probability of the walking passengers to stare at the same location as the stimulus group and found that the probability of gaze-following grows together with the stimulus group size (Milgram et al., 1969). However, as Gallup et al. (2012) point out, the observed behaviour by Milgram et al. (1969) might also be due to the quorum-like, instead of gaze-following, response.

The presumed reflexive bottom-up effect of gaze cues has come under scrutiny in recent research. For example, using eye-tracking Itier et al. (2007) studied the reflexive nature of gaze following with two different tasks requiring either explicit gaze or head orientation judgement. They also modified the task difficulty with miss-matching head and gaze orientations. The authors found that the participants made saccades towards the eye region also in the head orientation judgement task, but the eyes were not first and so often attended in head orientation judgement as in the gaze orientation judgement task, which suggests that gaze following is task dependent (Itier et al., 2007). Vecera and Rizzo (2006) also argue for a voluntary control of visuospatial social attention, which is based on their finding that although a patient with frontal-lobe damage could not follow gaze cues, he was still able to use peripheral cues to direct his attention.

Finally, a difference can be made between the concepts of shared gaze and joint visual attention, which are both widely appearing in the literature (Siposova & Carpenter, 2019). According to Frischen et al. (2007) shared gaze can be defined as a "spatial orientation response in the general direction of the observer's gaze", whereas joint visual attention means "orienting to the object that is the current focus of another's attention" (Frischen et al. 2007, p. 698). Hence, analysing JVA can be argued to require detecting mutual object-based levels of attention instead of mutual space-based attention alone. An exception is made by Nummenmaa and Calder (2009), who have combined the terms shared gaze and joint visual attention, defining joint visual attention as "sharing a common focus of attention such as an object or a spatial location with another individual". Another concept appearing in the literature is shared or joint attention, which is considered the highest state of mutual attention. It requires all interaction participants to be aware of each other's attentional state and, therefore, extends beyond sole visual attention (Fischen et al., 2007). In this study, joint visual attention (JVA) is defined similarly to Nummenmaa and Calder (2009), without making any space- or object-level differentiations.

2.3.1 Joint visual attention and memory

Although only few studies have explicitly investigated the effect of joint visual attention on individual long-term memory, research indicates that joint visual attention enhances both visual working memory (Gregory & Jackson, 2019, Gregory & Jackson, 2017, Gregory & Kessler, 2022) and to some extent, long-term memory (Kim & Mundy, 2012, Dodd et al. 2012). In their study, Gregory and Jackson (2017) combined the gaze cueing task with traditional WM task to study the effect of joint visual attention on visual working memory with different item loads (4, 6 or 8 coloured items to be remembered). The gaze cue was first presented either towards the items (valid), away from the items (invalid), or remained direct (uncued). The items appeared 500 ms after the gaze cue on either the validly cued or invalidly cued side and remained for 100 ms. They found that the working memory accuracy for the colours of the items was significantly better in validly cued conditions. Additionally, the authors researched the effect of arrow and motion cues but did not find any similar improvement in working memory for coloured items as they did with gaze cues (Gregory & Jackson, 2017).

In their second study, Gregory and Jackson (2019) focused on exploring the different mechanisms related to gaze cueing to better explain the effect of joint visual attention on working memory. According to the authors (2019), two different theories have been adopted to explain the effects of gaze cueing and following, and the resulting state of joint visual attention, on working memory. One is based on the social-attention account, where gaze cues are expected to enhance memory simply because visual attention is more directed towards the looked-at location. In addition, the effect of gaze cue on memory is found to diminish if the stimulus is presented for 1000 ms, compared to 250 ms and 500 ms, which supports the reflexive shift of visual attention towards the gaze cue (Dodd et al., 2012). Many gaze cueing tasks comparing gaze cues vs. non-social gaze cues have proved this theory of reflexively guided visual attention (Cole et al., 2015).

Another hypothesis regarding the gaze cueing effect on working memory is based on mental state and social cognition theories. According to this paradigm, the observer would not only redirect his/her gaze towards the location of the gaze cue but adopt the perspective of the gaze cuer. Hence, gaze following is expected to activate more detailed information processing about the object the gaze cuer is looking at (Gregory & Jackson, 2019). Neuroimaging studies have reinforced this theory of mind paradigm, demonstrating that the ventromedial prefrontal cortex, a key brain region activated in theory of mind tasks, is also engaged during gaze perception (Williams et al., 2005). Reversely, theory of mind tasks are found to engage the same regions of posterior superior temporal sulcus (pSTS) than in gaze perception tasks (Nummenmaa & Calder, 2009). Chuang and Hsu (2023) also made an intriguing discovery that even pseudo-mutual gaze can improve the efficiency of joint visual attention in (remote) collaborative work, which was shown in hyperscanning EEG data as interbrain synchronisation (IBS) coupling in frontal–central regions. These findings suggest a neural basis for the theory of mind processes activated in social interaction.

A common experiment to test whether the gaze cueing effect on working memory is caused by the social attention or mental state theory is to apply open and closed barriers to the gaze cueing task (Gregory & Jackson, 2019). In this setup, either an open or closed barrier is placed in between the gaze cue and the presented items. The barrier could also be something else than a physical (representation of a) wall, for example, gazers wearing either opaque or transparent mirrored goggles (Teufel et al., 2010). The key is to modify the experiment so that the observer needs to interpret to what extent the presented items are visible for the gaze cuer (Gregory & Jackson, 2019).

Research shows contradicting results about the effect of the closed barrier. For example, a closed barrier has been found to modulate object appraisal (Manera et al., 2014), but no effect has been found on reaction times and task accuracy (Cole et al., 2015). However, regarding visual working memory, Gregory and Jackson (2019) found that when the barriers were closed, the gaze cueing effect on visual working memory were abolished, which supports the mental state theory. In addition, Kim and Mundy (2012) argue that joint (visual) attention enhances memory through the depth of information processing. This enhancement is supported by the theory of parallel and distributed social-information processing, which encompasses processing both one's own visual attention and the visual attention of others. The depth of information processing has been proven by neuroimaging studies that examine the activation of different brain areas during social face-to-face interaction, and in between the conditions of joint visual attention and solo attention. For example, an fMRI study by Redcay et al. (2010) revealed that especially the right pSTS and right temporoparietal junction (TPJ) and their bilateral parts show greater activation in joint visual attention and live social interaction, compared to solo attention and recorded social interaction. These brain regions are generally regarded crucial for social perception, perceiving intentional actions (Saxe et al., 2004), understanding others' mental states and shifting attention (Krall et al., 2015).

Regarding the effect of gaze cueing on long-term memory, Kim and Mundy (2012) found that in virtual social interaction, gaze cueing as initiating joint attention (IJA) enhances recognition memory of pictures compared to gaze following as an act of responding to joint attention (RJA). The difference was present even when the total picture viewing time was the same. To explain the enhanced memory encoding in IJA condition, Kim and Mundy (2012) hypothesise the role of egocentric spatial information processing in relation to the referenced, allocentric spatial information processing, which is found to facilitate episodic memory retrieval (see Gomez et al., 2009). Other factors that they name is the volitional control related to IJA, which is shown to recruit neural reward circuits (see Schilbach et al., 2010) and the observed, more effortful attention control related to gaze following in infants, as presented by Vaughan Van Hecke et al. (2012) (Kim & Mundy, 2012).

From the social cognition point of view, joint (visual) attention can enhance memory based on the collective memory theory, which states that memory is not solely an individual process but also socially constructed within groups and societies. Although collaborative groups remember less than the combined memories of individuals, they do remember more compared to individual participants working alone, as shown in the study by Weldon and Bellinger (1997).

There is substantial evidence suggesting that gaze cueing, as well as prolonged engagement in joint (visual) attention, can enhance memory. However, most of the research on joint visual attention presented so far has been conducted in laboratory settings, and some contradictory results exist regarding the effects of joint visual attention in naturalistic, real-life social interactions (Kuhn et al., 2016, Caruana et al., 2017), as presented in the next section.

2.3.2 Real-life studies on joint visual attention

A common method to study joint visual attention in real-life social interaction is to apply mobile eye-tracking technology to analyse the amount of gaze coupling, which is then visualised as gaze cross-recurrence plots (CRPs) (Villamor & Rodrigo, 2019) or graphs (Richardson & Dale, 2005). For example, Schneider and Bryant (2024) studied the interplay between moments of collaboration and cooperation[1] of dyads by using mobile dual eye-trackers. Based on prior literature (see i.e., Richardson & Dale, 2005), they hypothesised a positive correlation between groups' productiveness and JVA. In addition, they also assumed that the more productive groups work in cycles of collaboration and cooperation. They tested the hypotheses by correlating the quantitative measures of JVA from the eye-tracking data with qualitative, subjective questionnaires that measured the quality of participant's collaboration, task performance and learning gains. Deviating from prior studies of JVA, they also measured the absence of JVA. They replicated the previous findings where joint visual attention positively correlated with collaboration quality. Moreover, they also found an increased correlation coefficient with collaboration quality by analysing the cycles of collaboration and cooperation integrating speech data and gaze cross-recurrence graphs (Schneider & Bryant, 2024).

Schneider and Bryant (2024) focused on the dynamics of joint visual attention in relation to collaboration cycles. In fact, a more detailed analysis of gaze cueing and following covers tracking the amount of initiating joint (visual) attention (IJA) and responding to joint (visual) attention (RJA), as presented by Kim and Mundy (2012). Especially considering the effect of JVA on learning, research has found that the metric of JVA as percentage of time spent mutually looking at certain AOI is not always the most indicative. Instead, in natural social interactions, it is preferable to analyse JVA in terms of balanced levels of collaboration (see Abdu & Schwarz, 2020), and to consider visual leadership roles in initiating JVA to recognize "free rider" or "partner dominance" effects (Schneider et al., 2018).

In face-to-face interaction studies, joint visual attention has been found to support learning (Schneider & Pea, 2013) and increase the quality of product (Fındık-Coşkunçay & Çakır, 2022), interaction (Jermann & Nüssli, 2012) and as mentioned, collaboration (Schneider et al., 2018, Schneider & Bryant, 2024). Deviating from other JVA studies that have focused on the increased or balanced interaction levels and subjective evaluations to evaluate the positive effect of JVA, Fındık-Coşkunçay and Çakır (2022) studied JVA during online collaborative business modelling and directly measured the quality of the outcome. They discovered that JVA, operationalized as gaze cross-recurrence, significantly predicted the overall quality of the collaboratively produced business process model, including its syntactic, semantic, and pragmatic aspects.

It has been suggested that in ecological settings the degree of joint attention in face-to-face interaction is dependent on the assumed level of interaction (Hessel et al., 2023, Gallup et al., 2012), which is presumably high in collaborative tasks. However, some researchers have found that people look at social stimuli less in real-life situations than expected, even in situations where social

---

[1] Cooperation and collaboration are regarded as two different modes of group work, collaboration referring to the shared attention and coordinated actions of the group, whereas cooperation refers to the individual actions of the group members without shared attention (Abdu & Schwarz, 2020).

interaction would be concurrent with social norms, such as while eating together (see i.e., Wu et al., 2013).

Other collaboration studies studying the so-called grounding or coordination costs of communication have often compared the effects of shared gaze and shared voice, or shared-gaze-plus-voice on collaboration. These studies have found that collaboration is often enhanced by shared gaze alone, at least in visual search tasks (see i.e., Brennan et al., 2008, Siirtola et al., 2019). This effect has been explained by the authors as substantial coordination costs related to the shared-gaze-plus-voice condition. Another study by Hessels et al. (2023) suggests that verbal communication does not significantly enhance collaboration performance. They studied shared gaze behaviour during a collaborative face-to-face Duplo-model copying task, where they manipulated the visibility of the blocks and whether verbal communication was allowed. Interestingly, they found that allowing verbal communication did not result in improved performance or alterations in gaze behaviour when all blocks were visible to both participants (Hessels et al., 2023).

However, some contradictory findings exist regarding the effects of shared gaze on collaboration. For example, McCarley et al. (2021) found that shared gaze did not improve the performance of virtually collaborating dyads in visual monitoring tasks when a cursor on the screen represented the other team member's gaze target. The authors suggest that the performance enhancing effect of displayed shared gaze is task dependent and can potentially cause redundancy in attentional scanning (McCarley et al., 2021), although the decrement effect of shared gaze could also be related to offline collaboration.

Finally, it must be highlighted that eye-tracking data only reveals the amount of overt joint visual attention. Hence, the key limitation of using eye-tracking technology to study joint attention include the fact that in real-life face-to-face interactions, gaze is only one type of communication signal and an indicator of joint attention among head-movements, facial expressions and verbal communication (Caruana et al., 2017, Brône & Oben, 2018).

To summarise, the degree of JVA in face-to-face interaction may vary depending on assumed levels of interaction. Contradictory findings exist regarding the effects of shared gaze on collaborative tasks, with some studies showing no improvement in performance, or even performance decrements. Although gaze following can be regarded as a reflexive act and crucial for a human intentionality detection system, eye-tracking data only captures overt visual attention, while real-life interactions involve multiple communication signals beyond gaze (Brône & Oben, 2018). However, based on previous JVA research in real-life social interactions, it can be hypothesised that JVA enhances the collaboration quality and long-term memory of the outcome.

2.4 Tactile artefacts used in collaboration

Besides examining the effect of (joint) visual attention on long-term memory, this thesis examines how tactile artefacts used in a collaborative context can enhance individual long-term memory. The underlying assumption is that tactile artefacts effectively direct joint visual attention, which is known to enhance working memory and, to a lesser extent, long-term memory. This chapter examines additional factors and relevant theories related to face-to-face interactions that involve the use of tactile artifacts to facilitate collaboration, and how these factors may influence memory.

2.4.1 Tactile artefacts as visual collaboration aids

Collaborating with tactile artefacts adds visual information elements to the communication. In general, it is believed that visual information is faster processed compared to textual information, although no explicit comparative studies exist (Tusher, 2022). However, it has been researched that adding visual information in the form of images or diagrams to support textual information often increases the comprehension of the message (Tusher, 2022).

Combining verbal and visual communication methods in collaborative strategy work became popular in the 1990s. Beyond visual communication, researchers became interested in applying cognitive approaches to study strategic management. A new concept of ʻstrategic mapping' was introduced in a book titled *Mapping Strategic Thought* (1990) that provided a set of cognitive mapping methods for strategic management. According to the pioneers of strategic mapping, a map is useful as "a graphic representation that provides a frame of reference" (Fiol & Huff, 1992). In the business context, cognitive maps were seen as interesting because managers use these ʻrepresentations of thought' to graphically display the firm's current strategic position. Moreover, the cognitive maps of managers were regarded as action oriented as they held "the promise of identifying alternative routes to improving that position" (Fiol & Huff, 1992) —in case they were effectively displayed for others.

Compared to traditional diagrams and mind maps, tactile artefacts provide an effective method for strategic mapping practices because objects can both embody knowledge and facilitate the creation of new knowledge (Teh et al., 2021). For example, assigning symbolic meanings to objects and describing a strategy with these symbolized objects can trigger creative thinking, help users explore abstract concepts and uncover the root causes of a problem. Teh et al. (2021) have studied and described the LEGO® SERIOUS PLAY® (LSP) builds as ʻco-created knowledge objects' that result and lead to knowledge integration. Similarly to Entwistle & Marton (1994), Teh et al. (2021) define knowledge objects analogous to cognitive schemata (see Neisser, 1967), which as mental templates help to organize and understand new information. Based on their study about the role of LSP bricks in preparing for a debate competition, Teh et al. (2021) suggest a phenomenological model based on personal and shared schemata to describe the dynamic functions of co-created knowledge objects. In their model, personal schemata refer to the ideas and cognitive structures of an individual which is related to the construction of knowledge objects, whereas shared schemata are KOs accessible in the social realm (Teh et al., 2021, p. 558). For example, an individual might use their personal understanding of debate strategies to build a specific knowledge object with LSP bricks (personal schemata). However, a different set of LSP bricks might be used by the debate team to construct a shared model of their debate strategy, which everyone in the team can interpret and use (shared schemata).

Studies on communication and collaboration have also suggested that the effective coordination of dyads is dependent on shared visual information (Gergle et al., 2013). Similar to the visual feedback a chef gets when preparing a meal by observing the cooking process, tactile artefacts also serve as a hands-on reference for the current state of the collaborative task. According to Gergle et al. (2013) shared visual information and immediate visual feedback affect two coordination processes: situation awareness and conversational grounding.

2.4.2 Tactile artefacts and memory

The multicomponent model of working memory by Baddeley and Hitch (1974) states that working memory is divided into separate stores of visuospatial, verbal and auditory information (Styles 2005, p. 8). In fact, neuroimaging studies have proved the existence of different domain-specific cortical networks for verbal and visuospatial working memory (see i.e., Gruber & Cramon, 2003). In addition, behavioural studies have revealed that participants tend to perform better when the two tasks engage different types of working memory stores (Baddeley & Warrington, 1973). This observation is consistent with the principles of the dual coding theory proposed by Allan Paivio (2013). According to the dual coding theory, visual information is often implicitly named verbally, which facilitates its encoding into long-term memory. Therefore, tasks that involve both visual and verbal processing are more likely to engage separate memory systems, leading to enhanced performance in dual-tasks scenarios (Baddeley & Warrington, 1973). As collaborating with tactile artefacts entails an explicit verbal naming and referencing of the artefacts, they can be argued to effectively reinforce both working memory and long-term memory.

In addition to regarding tactile artifacts as visual information elements that can enhance memory, their use within a narrative context also has a theoretical basis for potential memory improvement. The narrative hypothesis suggests that individuals perceive and communicate experiences through stories or narratives, which can be seen as an extension of schema theory. According to the narrative hypothesis, people understand and make sense of events through storytelling processes (León, 2016). In contrast, schema theory posits that experiences are interpreted within the framework of pre-existing mental schemas (Neisser, 1967). While Neisser (1967) emphasizes that these mental schemas determine what is stored in memory, León (2016) argues that narrative memory—considered a subset of episodic and semantic memory—stores information that has narrative characteristics. This subset of memory can be seen to cover the memory for the meaning of the tactile artefacts used in face-to-face collaboration, which is different from the visual memory for the tactile artefacts.

Recent neuroscientific research has examined the neural structures of encoding and retrieving a narrative. An fMRI study on schematic knowledge encoding and retrieval reveals that the memory for narrative details (event schemas) is supported by distinct brain networks during the encoding and retrieval of the story (Masís-Obando et al., 2022). Masís-Obando et al. (2022) hypothesise that the distinctiveness of the encoding and retrieval phases of the story might be explained through the theories of schematic scaffolding and sequential memory cueing strategy. During the encoding phase of the narrative, "a structured set of attachment points" for the details of the story are built, which is shown as an activation in the medial prefrontal cortex (mPFC) (p. 12-13). Instead, as the retrieving of the story do not show as strong activation in the mPFC, the sequential memory cueing strategy might support the retrieval process of the narrative by first memorising the events of the story, and then its related details (Masís-Obando et al., 2022, p. 12). The study by Masís-Obando et al. (2022) showed that the schema representations in visual cortex contribute to memory, but the networks differ similarly during the encoding and retrieval phases. In the study, the activation in the proportions of the left visual cortex had stronger relationship to memory support during encoding than during retrieval, whereas the bilateral visual cortex had stronger relationship to memory support during the retrieval process (p. 13). The activation of the visual cortex during both the encoding and retrieval processes supports the hypothesis that using visual communication aids can facilitate the formation and activation of schemas, thereby enhancing narrative memory. However, while the use of tactile artifacts in a narrative context has a strong theoretical basis for enhancing memory encoding and retrieval, it is difficult to predict whether their use will improve both narrative and visual long-term memory in a collaborative context.

2.5 Research gaps

The current study aims to address some of the research gaps identified in previous studies on JVA, mobile eye-tracking, and long-term memory. Firstly, most research on JVA has been conducted in laboratory settings, with real-life studies primarily focusing on JVA behavior in autistic children (e.g., Mundy, 2018). Furthermore, research on JVA in real-life interactions has often considered only dyadic interactions, while most real-life collaboration occurs in larger groups. Therefore, further exploration is needed to analyse the dynamics of JVA, including gaze initiation and following, in real-life interactions and their impact on collaboration and communication. Additionally, studying the effects of JVA on long-term memory in real-life interactions is valuable not only for improving business practices but also for developing new methods to enhance learning, and contributing to the expanding field studying the effects of JVA on collaborative learning (e.g., Zhang & Barmaki, 2020).

Additionally, while mobile eye-tracking technologies have become more accessible in recent years, the analysis methods have primarily relied on visualizations like gaze cross-recurrence graphs. These visualizations are not suitable for statistical analysis as they do not capture the temporal dynamics of JVA, such as the frequency or duration of fixations during JVA. However, mobile eye-tracking has the potential for analysing face-to-face collaboration dynamics and their effects by tracking visual attention in real time. Finally, while incorporating visual communication aids into collaboration and simultaneously encouraging the use of narratives has a strong theoretical basis for enhancing learning and memory, this has not been extensively researched in real-life interactions.

3 METHODS

3.1 Research questions

The main research question is “If and how does (joint) visual attention to tactile artefacts in face-to-face collaboration enhance individual long-term memory of the group’s outcome?”. Based on the literature review, it is hypothesised that joint visual attention enhances individual long-term memory of the group’s outcome. Although joint visual attention is the main interest of this study, also individual visual attention to tactile artefacts will be studied.

Other research questions are related to different types of memory because the group’s outcome can be understood in two different, but related ways. The participants can either have a strong visual memory of the outcome in terms of the artefacts used and their assigned meanings, or a strong narrative memory of the outcome—referring to the jointly built common understanding to the topic with the help of the artefacts. Therefore, the second research question is “If and how does (joint) visual attention enhance visual and/or narrative long-term memory of group’s outcome?”. It is hypothesised that JVA enhances both visual and narrative long-term memory of the final build because the artefacts are seen to ‘embody’ the details of the outcome. However, in the light of neuroimaging studies, if the retrieval of the outcome activates sequential memory cueing strategy as recalling a story through its events (see i.e., Masís-Obando et al., 2022), the artefacts as details of the outcome might not be as well remembered.

3.2 Hypothesis

H1. (Joint) visual attention to tactile artefacts enhances narrative long-term memory of the outcome.

H2. (Joint) visual attention to tactile artefacts enhances visual long-term memory of the outcome.

H3. Initiating joint visual attention to tactile artefacts enhances visual long-term memory of the outcome.

H4. Initiating joint visual attention to tactile artefacts enhances narrative long-term memory of the outcome.

3.3 Experimental design

The study design is nested, where 8 groups of three to four participants used stand-alone 3D objects in a facilitated workshop. The groups were recruited from different companies and organisations within Helsinki region, Finland. A total of 29 participants took part in the study, consisting of 20 males and 9 females, with a mean age of 45.2 years (SD = 9.96). The only prerequisite for participation was that the company or organisation deals with customers, as the common task of the workshops was to develop a strategy to enhance customer engagement. No compensation was provided for participating in the study.

Before the workshop, the participants were introduced to the study by two experimenters—one of them also acting as the workshop facilitator. First, the participants signed the informed consent form and filled in the demographics questionnaire, which included questions about their areas of work and expertise in the workshop language (Finnish or English). Next, the participants were instructed to wear the Pupil Invisible eye-tracking glasses developed by Pupil Labs, with the recording Android One Plus 8 -smartphones attached to their arms with an armband. If necessary, participants' own prescription glasses were replaced by adjusting the lenses of the Pupil Invisible eye-tracking glasses accordingly. The temporal synchronisation of the four eye-tracking glasses was ensured by connecting them to the same router with an internet connection and using the time provided by the network. This synchronisation method ensured that the time difference between the devices was within 20 milliseconds (Pupil Labs, 2024).

With the eye-tracking glasses on, each participant was seated in front of a table that had a standard sized validation poster by glassesValidator placed horizontally (Niehorster et al., 2023). The glassesValidator is an open-source program for automatic quality detection of the eye-tracking recording accuracy and precision (Niehorster et al., 2023). First, the participants were asked to look at the red mark on the top left corner of the poster while the experimenter manually calibrated the glasses with the Offset Correction feature on the Pupil Invisible Companion application. The distance of the poster from the participant’s eyes (approximately 40 cm) simulated equal distance as the placement of most of the tactile artefacts used during the workshop. After the Offset Correction, participants performed the validation task similarly as described in Niehorster et al.

(2023) by sequentially looking at each of the marks in the poster for one second, in a descending reading order. As a data quality guarantee, they were asked to perform the sequential looking twice.

After the validation procedure the participants were seated in a room in front of a table that had a joint building board with 64 pieces of tactile artefacts on it (see Figure 2). To spatially align the eye-tracking data, AprilTag (tag36h11 family) fiducial markers (Olson, 2011) were attached to a wooden board, which was then placed on top of the table. Before the workshop, participants were instructed to avoid placing artefacts on top of the AprilTag markers. The room also entailed a screen where instructions of the facilitated workshop appeared (also marked with AprilTags), a video camera that recorded the social interaction during the workshop, and an extra light.

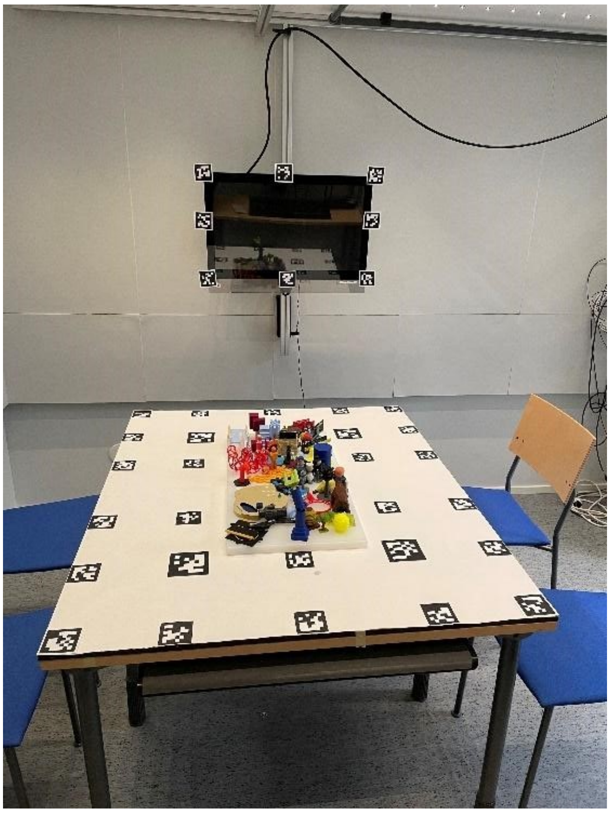


Figure 2. Experiment room.

The workshop was divided into three main phases, consisting of a warmup task, an individual task and a group work task. A 10-minute break was given in the middle of the group task. In order to detect the different phases from the eye-tracking data, participants were instructed to look at the screen at the beginning and at the end of each phase, after which a black screen was rapidly displayed to them. The black screens were manually detected and annotated as events from the Pupil Cloud eye-tracking scene video stream.

The main task of the workshop was to build a common understanding of how to enhance customer engagement by using tactile artefacts. As an outcome, the board entailed a final build consisting of tactile artefacts, which were iteratively chosen, assigned meaning and ordered during the workshop. The group work task ended after one of the participants described the meaning of the final build (also referred to as the group's outcome), which was recorded with a phone by another participant. The recording phone was provided by the experimenters to prevent participants from rewatching the video before the surprise follow-up interviews held 3-4 weeks after the workshop, which consisted of questions about their memory of the group's outcome (see Appendix 1).

During the workshop, after both the individual task and the group work task, participants were asked to fill in a Flow Short Scale (FSS) and a NASA Task Load Index (TLX) questionnaires. After the group work task, in addition to the FSS and NASA-TLX questionnaires, a Five-Dimensional Curiosity Scale Revised (5DCR) was added, but without applying reverse-coding for the stress tolerance -questions (Kashdan et al., 2020). In addition, after the group work task seven items were included that were related to the group work dynamics, equal participation and satisfaction to the outcome. However, the responses to the questionnaires were not assessed in this thesis.

After the workshop, all participants repeated the glassesValidator validation task (Niehorster et al. 2023) to detect discrepancies in the eye-tracking data accuracy and precision. After the eye-tracking glasses were taken off, the participants gathered for a short reflection discussion with the workshop's facilitator.

Three weeks after the workshop the participants received an email invitation to participate in an online, structured follow-up interview, which was organised within a week from the invitation. All interviews were recorded. The follow-up interview was divided into two parts (see Appendix 1). The first part consisted of free recall questions about the memory of the group's outcome. The participants were first asked to memorise the artefacts in the final build and their assigned meanings, which allowed studying the visual long-term memory of the group's outcome. After memorising the individual artefacts, the participants were asked to describe the meaning of the final build in their own words, as if they would tell the outcome to a third person who did not participate in the workshop. This section of the follow-up interview was used to analyse the narrative long-term memory of the group's outcome. The second part of the follow-up interview was a cued recall, where a picture of the final build was shown to the participants and the same questions were repeated.

3.4 Data analysis

3.4.1 Joint visual attention

In the Pupil Cloud platform, ten areas-of-interest (AOIs) were first defined according to the detected AprilTag markers (Olson, 2011) with the Marker Mapper Enrichment feature. Nine of the AOIs were segmented from the table and the screen was designated as one AOI. The joint building board was marked as a separate AOI, used exclusively in this study. The Marker Mapper feature returns the fixation coordinates in 2D surface coordinates when at least two AprilTag markers of the defined AOI are detected. The fixation position (x and y) detected inside of each AOI were returned between the coordinates of 0 (top left) and 1 (bottom right).

Previous eye-tracking research has used a computer vision technique called homography estimation to detect joint visual attention (see i.e., Schneider, 2020). Homography estimation is frequently employed to establish the relationship between two images captured from different viewpoints. However, conducting homography estimation from videos demands a substantial amount of computational power. Fortunately, the experimental design used in this study allowed a different approach to compute joint visual attention as the participants were seated during the workshop. Hence, instead of using homography, joint visual attention was calculated from the detected fixations within the AOIs by flipping 180 degrees of the 2D coordinates for two of the participants, who were sitting on the other side of the table. After the flipping of the coordinates of the two participants, all fixations under 50 ms and above 2000 ms were filtered out.

Joint visual attention was computed with a Python script developed together with Aalto University Research Software Engineers. We first calculated pairwise matrices to analyse overlapping visits from the pre-processed visit data. The pre-processed visit data entailed all unique gaze visits, which were defined as subsequent fixations on the AOI. A distance matrix represented Euclidean distance between the visit positions of each pair of participants, where the mean visit position was calculated from the detected fixations during a visit. As the participants' visits cannot be expected to occur simultaneously in a millisecond and millimetre level, time and proximity windows were given to detect joint visual attention events. The distance matrix was used to calculate a Boolean matrix based on a threshold of 5 cm between the visit positions. The proximity threshold is dependent on the experimental design, research interest, as well as the accuracy and precision of the mobile eye-trackers. Based on the calibration procedure and validation results acquired with glassesValidator, the mean accuracy of the eye-trackers was discovered to reach 2,98° before the workshop and 3,09° after the workshop (see all results in Section 4.1.1). Considering that the participants looked at the table from a maximum distance of one meter, the accuracy offset of 3° of the calibrated eye-trackers equals approximately 5.24 x 5.24 cm area on the joint building board. Another measure used to define the proximity threshold was the average size of the artefacts and the size of the joint building board (66 x 33 cm), where the artefacts were placed during the group work task of the workshop.

Similarly, a Boolean matrix of time proximity was calculated based on the differences between the start and end times of the visits. A time proximity window of ±2 seconds was applied to each visit's start and end timestamps, based on research indicating that listeners' eye movements lag behind speakers' eye movements by approximately 2 seconds (Richardsson & Dale, 2005).

From the distance and time proximity matrices, we created an adjacency matrix to construct a graph of events, which described whether two visits were connected into the same joint visual attention event, considering both the spatial and temporal proximity of the visits. Besides analysing the number of joint visual attention events and their duration (ms), the initiators and followers of JVA events were detected to investigate the group's interaction dynamics in more detail. Based on prior research, it was assumed that the dynamics of JVA could be a better metric instead of JVA alone. As covered in the literature review, initiating joint (visual) attention (IJA) is found to enhance recognition memory of pictures compared to gaze following, which can be a relevant metric when studying the visual long-term memory of the group's outcome (Kim & Mundy, 2012). Therefore, for each joint attention event, gaze initiator and followers were assigned. In addition, since joint visual attention can also be understood and measured in degrees or as a scale of jointness (see i.e., Siposova & Carpenter, 2019), the number of participants involved in joint visual attention was determined.

The full code used to analyse joint visual attention with mobile eye-tracking and AprilTags is available on GitHub: https://github.com/Janinka-1/thesis-JVA

3.4.2 Long-term memory

To research the long-term memory of the group's outcome, the follow-up structured interviews and the workshop outcome videos were transcribed from speech to text with Aalto speech2text speech recognition app, which is developed and performed by the Aalto University School of Science "Science-IT" project. The follow-up interviews consisted of both free and cued recall questions, but only the answers from the free recall questions were extracted for further analysis in this study. However, the long-term memory performance itself was operationalized and analysed in two different ways. First, since the initial part of the interview consisted of questions related to individual artefacts and their assigned meanings, the memory of the group's outcome was analysed as the proportion of remembered artefacts and their assigned meanings in the final build. This analysis method is considered to indicate the activation of visual long-term memory of the group's outcome, although it also entails the verbal referencing of the artefacts.

The second type of analysis is related to the part of the long-term memory interview where the participants were asked to memorise the meaning of the final build and describe it in their own words as accurately as possible. To address this part of the interview, coherent sentences were extracted both from the transcribed long-term interviews and from the final video to compare their semantic similarity. This method primarily researches the memory of the group's outcome through the activation of narrative memory, which is similar with verbal recall analysis (see e.g., Masís-Obando et al., 2022, p. 18). However, instead of manually coding each participant's memory performance (see e.g., Masís-Obando et al., 2022), the semantic comparison of each sentence's similarity with the final video was performed computationally using a sentence-transformer, which is a type of natural language processing (NLP) model. NLP models approach human language as a complex system that has a series of semantic and syntactic rules, which are related to the meanings of sentences and grammar, respectively (Kapetanios et al., 2013).

A key challenge in NLP is commonly either related to semantics or compositionality (Kapetanios et al., 2013). Compositionality refers to the structure of the sentences, meaning that the whole sentence should be a "systematic function of the meaning of its components" (Kapetanios et al. 2013, p. 133-134). This challenge is particularly relevant for complex language nuances that don't always follow straightforward rules (Kapetanios et al., 2013, Rothman, 2021). To tackle this challenge, modern NLP has shifted from rule-based methods to context-aware approaches (Kapetanios et al., 2013). Current NLP models, trained on extensive text data, use statistical and machine learning principles to predict the next word by analysing semantic and syntactic patterns. This approach emphasises statistical and machine learning principles over formal linguistic theories. A pivotal aspect of their functionality lies in the representation of words as vectors within a multidimensional space—an approach commonly referred to as word-embedding, which can capture the semantic relationships between words based on their contextual usage (Kapetanios et al., 2013). Contextualized word embeddings are created using transformer models, which learn language representations from large text corpora. Recent bidirectional transformers, which analyse both left and right contexts, provide better contextual embeddings compared to earlier unidirectional models (Rothman, 2021).

The transformer-architectures used in today's language-transformer models are based on deep learning model architecture, which was introduced by Vaswani et al. (2017) in a famous article titled "Attention is All You Need". Those transformer-architectures differ from recurrent neural network models so that they do not rely on sequential processing, but instead they use attention mechanisms to consider all positions in the input sequence simultaneously (Rothman, 2021). When the recurrence is abandoned, the transformers are better at handling long-range dependencies with different attention mechanisms. The self-attention sequence-to-sequence operations are similar to human attention in reading tasks; they look at each word and decide how much attention will be directed to the word. In case the word is important for comprehension, more focus is given by assigning weights to the word. In practice, the self-attention mechanism counts the dependency of each word in a sequence (Rothman, 2021, p. 5).

Another key feature of the transformer-architecture is an encoder-decoder structure. The encoder-decoder structure means that the encoder processes the input sequence whereas the decoder regenerates the output sequence. The original encoder structure presented by Vaswani et al. (2017) entails six layers, each of them containing two main sub-layers: a multi-headed attention mechanism and a feedforward network (Rothman, 2021, p. 6). The multi-headed attention performs similar functions through each layer, looking at different associations of words, and each layer learns from the previous one (Rothman, 2021, p. 8).

After the pre-training, language-transformer models can be fine-tuned for several specific tasks. For example, they can be used to detect semantic textual similarity, which is an NLP task comparing the degree of semantic similarity in between two sentences, assigning them a similarity score between 0 and 1. The semantic textual similarity has been widely used in NLP tasks, such as in information retrieval, question answering and summarization (Rothman, 2021).

To capture the semantic textual similarity between the long-term memory interview answers and the final story of the group's outcome, an XML-RoBERTa based cross-lingual English-Finnish Sentence-BERT (SBERT) model was employed as a sentence-transformer. The model is developed by Mikko Moisio (2021) and it relies on the work by Reimers et al. (2019) to generate semantically meaningful sentence embeddings (Reimers & Gurevych, 2019). The model has a good performance in both Finnish and English, and it also outperforms the FinBERT model (Moisio, 2021, p. 30-31), which was the sole alternative to use in this study. Using the sentence-transformer was seen as less biassed to evaluate the memory performance of the participants than the commonly used subjective coding. However, for the interval validity, the obtained results were manually screened and validated by randomly choosing one participant's data from each group.

The model was applied to compare the cosine similarity between sentence-transformer embeddings of the sequences in the final video (original sentences), and with the embedding sequences of each long-term memory interview answer (comparison sentences). A threshold for semantic similarity between the sequences was set at 0.7, meaning that the sentence pairs below the threshold were regarded as dissimilar. The threshold was determined after manually screening and evaluating the model's accuracy to detect similar contexts from the sentences.

In cases where multiple comparison sentences exceeded the threshold, only the closest pair with the original sentence was considered to avoid duplication. Therefore, the total number of similar sentences could not exceed the number of sentences in the final video. However, it could exceed the number of sentences in the long-term memory interview answers. This logic allowed each comparison sentence to be used more than once, while ensuring that the original sentences were matched only once. This also ensured that no data from the interviews was missed, as individual sentences in the participants' responses might have included multiple details or events from the final story of the build. The results were presented as a proportion of similar sentences found in each long-term memory interview answer relative to the final video.

3.4.3 Statistical analysis

To test the relationship between JVA or initiating JVA and long-term memory, a binary logistic mixed-effects model was fitted the lme4 package in R (Bates et al., 2015) with visits of JVA or initiating JVA as a dependent variable. The binary logistic mixed-effect regression model is a type of generalised linear mixed model (GLMM) and an extension of the generalised linear model (GLM) (Demidenko, 2013, p. 331). In the binary logistic mixed-effects regression model, the response variable is binary (coded as 0 or 1) (Demidenko, 2013), which is applicable for the variables of JVA and initiating JVA. The visits of JVA or initiating JVA were chosen as the dependent variables, because the long-term memory answers only had one observation per participant and therefore, it could not be used as a dependent variable in a mixed-effects regression model. To test the effect of sole visual attention on long-term memory, a linear mixed-effects model was fitted with the visit duration as a dependent variable. Hence, although the statistical analysis used might seem counterintuitive, using visits as the unit of analysis allows for capturing variability and nuances at a more detailed level, providing a more precise understanding of the temporal dynamics and fluctuations in visual attention, JVA, and the initiation of JVA.

Hence, JVA and initiating JVA are the response variables Y, where 1 indicates a presence of JVA or initiating JVA and 0 represents their absence. As the response variable Y represents probabilities, the link function used in binomial regression is typically a logit function (also called log-odds function) (Roback & Legler, n.d.). The logit link function transforms the linear predictor (a combination of predictor variables) into the log-odds of presence:

$$logit(p) = log\left(\frac{p}{1-p}\right) = X\beta$$

In the function, p is the probability of presence, X are the predictor variables, and β are the coefficients. The coefficients β estimated from a binary regression model represent the effect of the predictor variables on the log-odds of the outcome. Exponentiating the coefficients (inverse of the logit function, known as odds ratios, OR) provides insights into how the predictors affect the odds of the binary outcome (Roback & Legler, n.d.).

Following the similar method to analyse the sole effect of visual attention on visual and narrative long-term memory, two linear mixed-effects (LME) models were fitted with lme4 package in R (Bates et al., 2015), using visit duration as a dependent variable. The LME model allows for modelling both fixed and random effects, and since the random effects account for variability across individuals or groups, it is well suited for analysing data with a nested design (Demidenko, 2013).

4 RESULTS

4.1 Data quality

4.1.1 Eye-tracking data

Eye-tracking quality is often reported in its accuracy, precision and data loss (Holmqvist et al., 2012). Accuracy describes the distance between reported gaze position and actual gaze position, whereas precision refers to the sample-to-sample root mean square (RMS) and describes the consistency of the calculated fixation points (Holmqvist et al., 2012). The standard deviation (SD) measures the dispersion of gaze points around the mean accuracy. The accuracy and precision of the Pupil Invisible eye-tracking glasses were measured with the glassesValidator procedure (Niehorster et al., 2023) before (B) and after (A) the workshop. The obtained results of mean accuracy (°), precision and standard deviation are presented in Table 1.

Participant ID	Accuracy (°) (B)	Precision (B) (RMS)	SD (B)	Accuracy (°) (A)	Precision (A) (RMS)	SD (B)
G1P1	4.98	1.26	1.61	3.83	0.84	1.01
G1P2	2.27	0.51	0.67	2.92	0.52	0.72
G1P3	2.62	0.59	0.72	1.83	0.33	0.72
G1P4	2.16	0.24	0.33	2.8	0.19	0.33
G2P1	2.26	0.35	0.57	2.8	0.45	0.72
G2P2	2.60	0.36	0.51	4.0	0.19	0.57
G2P3	1.94	0.23	0.36	1.83	0.28	0.49
G2P4	4.19	0.31	0.51	4.8	0.38	0.68
G3P1	2.09	0.49	0.73	2.9	0.42	0.49
G3P2	3.45	0.48	0.62	4.74	0.73	0.97
G3P3	6.76	0.36	0.49	4.63	0.46	0.7

G3P4	3.34	0.34	0.55	4.67	0.39	0.68
G4P1	1.49	0.33	0.68	1.69	0.34	0.64
G4P2	*NA*	*NA*	*NA*	6.61	0.56	0.98
G4P3	2.3	0.28	0.39	2.82	0.36	0.48
G4P4	4.47	0.3	0.42	3.43	0.43	0.59
G5P1	1.99	0.6	0.81	2.97	0.49	0.76
G5P2	2.46	0.47	0.67	1.89	0.5	0.64
G5P3	2.96	0.46	0.62	2.05	0.28	0.47
G5P4	3.46	0.23	0.36	1.89	0.328	0.49
G6P1	2.69	0.26	0.4	3.66	0.31	0.46
G6P2	2.71	0.32	0.52	2.78	0.49	0.72
G6P3	*NA*	*NA*	*NA*	2.53	0.75	1.05
G7P1	*NA*	*NA*	*NA*	2.2	0.36	0.53
G7P2	4.0	0.89	1.1	*NA*	*NA*	*NA*
G7P3	1.12	0.5	0.63	1.12	0.36	0.5
G8P1	4.36	0.27	0.44	3.18	0.39	0.62
G8P2	2.72	0.38	0.62	1.98	0.34	0.53
G8P3	2.04	0.35	0.55	3.88	0.33	0.53
Mean	**2.98**	**0.43**	**0.61**	**3.09**	**0.42**	**0.65**

Table 1. glassesValidator results of the eye-tracking data accuracy and precision measured before (B) and after (A) the workshop.

No standardised guideline exists to evaluate whether the accuracy is good enough because it is dependent on the research questions and study design. During the pilot studies, it was discovered that the uncalibrated Pupil Invisible eye-tracking glasses, which have a marketed accuracy of 4.6°, did not provide sufficient accuracy for this study. Therefore, the Offset Correction feature from the Pupil Invisible Companion App was used to achieve higher accuracies. The mean accuracy of the calibrated eye-trackers measured before the workshop was 2.98° (SD = 0.61) and 3.09° (SD = 0.65) after the workshop. To further analyse whether the accuracy was sufficient for the purposes of this study, the percentage of visit duration in each AOI was measured and visibly inspected (see Figure 3). Hypothetically, poorer accuracy could be reflected in the time spent looking at each AOI. The correlation between the average accuracy and visit duration percentage on AOIs was -0.23 (Spearman's ρ, p = .30) before the workshop, and -0.54 (Spearman's ρ, p < .01) after the workshop. The correlation between the mean accuracy measured before and after the workshop and visit duration percentage on the AOIs was -0.52 (Spearman's ρ, p < .01), indicating that better accuracy (lower value) had a statistically significant and moderate correlation to the visit duration percentage on the AOIs. The negative correlation means that as accuracy improves, visit duration percentage on AOIs tends to increase.

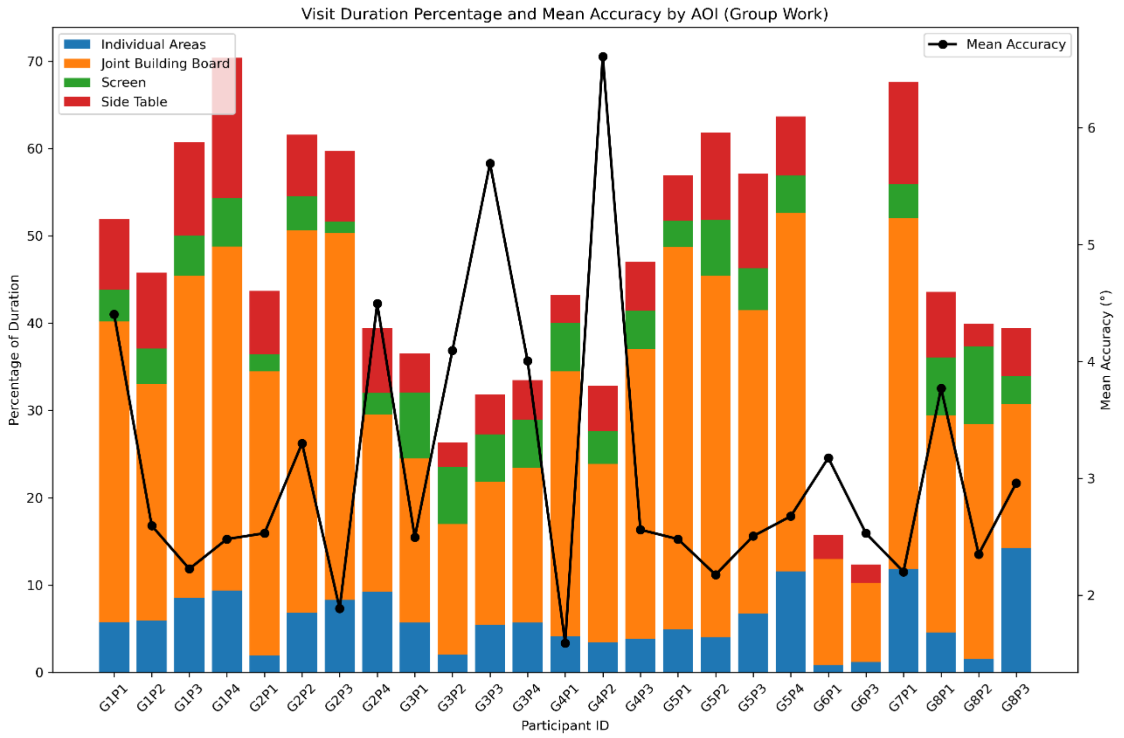


Figure 3. The distributions of visits on the AOIs during the group work task and the mean detected accuracy of the Pupil Invisible eye-tracking glasses.

4.1.2 Obtained data from the study

A total of eight groups participated in the study. Eye-tracking data from four participants (G4P4, G6P2, G7P2 and G7P3) were fully or partially missing due to recording errors during the workshop and therefore these participants' data were excluded from the analysis.[2] This means that the group number seven was also excluded from the analysis as no joint visual attention could not be detected with only one successful eye-tracking recording (see Table 2). The total number of participants having both successfully recorded eye-tracking recording data and participation in the long-term memory interview was 18, and they were divided in 7 groups.

Table 2. Obtained data from the study.

Group number	Recorded eye-tracking data during the workshop	Long-term memory interviews
Group 1	4/4	3/4
Group 2	4/4	3/4
Group 3	4/4	3/4

---

[2] The number after the letter G refers to the group number, while the number after the letter P refers to the randomized participant in the group.

Group 4	3/4	1/4
Group 5	4/4	4/4
Group 6	2/3	2/3
Group 7	1/3	3/3
Group 8	3/3	3/3
Total	25/29	22/29

4.2 Descriptive statistics

The analysed data comprised a total of 18 participants, including both female (*N* = 7) and male (*N* = 11). The average age of the participants was 45.4 years (SD = 10.2). In nested data continuous variables can be described by taking into account the structure and average the results by the clusters or combining the observations and averaging the variables (Wiley & Wiley, 2019). In this study with a nested design, the participants (*N* = 18) belong to the first level in the dataset. The second level is the group (*N* = 7) where the participants belong to. In Table 3, the group level variables were averaged by the groups, and the individual variables were averaged by the participants. The descriptive eye-tracking metrics concern the joint building board as no other AOIs were analysed in this study. As the variable of joint visual attention is binary, the visits were presented as the mean proportion of visits in JVA.

Level	**Key variables**	**Mean (SD)**
Group	Number of artefacts used	14.5 (5.6)
	Group work task duration (min)	52.2 (20.4)
Individual	Visits of JVA (%)	29
	Number of gaze following	74.1 (43.6)
	Number of initiating JVA	50.3 (32.2)
	Mean visit duration (ms)	2018.5 (4036.5)
	Mean number of fixations per visit	5.9 (10.4)
	Mean fixation duration per visit (ms)	337.1 (232.5)

Table 3. Levels of the dataset and key variables.

To validate the given time (±2 seconds) and proximity (5 cm) thresholds to determine JVA events, the maximum physical and temporal distance between the visits were calculated. The maximum physical distance between the visits during an event was 41 cm and the maximum JVA event duration was 110.6 seconds. The mean of the maximum distance between visits during a JVA event on the joint building board was 9.3 cm (SD = 5.1). The mean JVA event duration on the joint building board was 10.3 seconds (SD = 13.4), while the median JVA event duration was 5.6 seconds.

4.3 Data preprocessing and covariate selection

4.3.1 Selecting fixed and random effects for the models

Mixed-effects models can be built either iteratively, adding variables to the model, or adding all variables to the model and removing them one by one, to see how well they explain the variance at individual and group level (Demidenko, 2013). In this study, the fixed effects tested in the models were chosen beforehand based on the literature review. As the memory of the group's outcome was operationalized in two different ways, two binary logistic mixed-effects regression models were fitted to predict JVA and initiating JVA. Additionally, two linear mixed-effects models were fitted to predict the sole effect of visual attention on visual and narrative long-term memory. The visit duration was chosen as the dependent variable to study visual attention, as it indirectly encompasses both fixations and saccades, reflecting the total time spent looking at the joint building board.

The dependent variable for the binary logistic mixed-effects regression models is either JVA or initiating JVA. JVA is coded as 1 in case when the visits of at least two different participants are temporally and physically overlapping, as defined in section 3.4.1. Initiating JVA is coded as 1 in case the participant started the JVA event. The main continuous fixed effects include proportion of artefacts remembered (%), semantic similarity percentage (%), visit duration (ms), number of fixations during a visit (ms), group work task duration (min) and the number of artefacts in the final build. The mean fixation duration during a visit was excluded from the models due to its covariance with the number of fixations. The number of fixations was preferred as it is shown to affect the recognition memory (Tatler et al., 2005), and the attentional metric of visit duration indicates the total amount of visual attention directed towards the artefacts. The only categorical fixed effect included in the models is gender, to test whether the previously found gender differences on judging non-verbal cues are replicable (Hall, 1978).

For the linear mixed-effects model, the dependent variable of visual attention operationalized as visit duration is predicted by either the proportion of artefacts remembered (%) or the semantic similarity percentage (%). In addition, the number of artefacts used in the final build is included as a continuous predictor to assess its effect on directing visual attention (Table 4).

The chosen random effects for the models included unique visits, participants, and groups. Including participants and groups as random effects accounts for the nested structure of the study. The random effect for participants allows the findings to be generalized across different individuals, capturing the variability between participants. The group-level random effect accounts for variability across different groups, helping to understand how much of the variation is due to differences between groups. Including unique visits (detected continuous gaze on the AOI) as a random effect accounts for variability at the level of each visit, although it also entails a risk of overfitting by attempting to explain variance at a very granular level. However, designing the maximal random effects structure that is justified by the study design allows for the most reliable generalization of the results (Barr et al., 2013).

Model	**Dependent variable**	**Continuous fixed effects**	**Categorical fixed effects**	**Random effects**
1. Visual long-term memory	1.1 JVA 1.2 Initiating JVA	Proportion of artefacts remembered (%) Visit duration (ms) Number of fixations (N) Group work duration (min) Number of artefacts used in the final build (N)	Gender (Female/Male)	Participant Unique visit Group
2. Narrative long-term memory	2.1 JVA 2.2 Initiating JVA	Semantic similarity percentage (%) Visit duration (ms) Number of fixations (N) Group work duration (min) Number of artefacts used in the final build (N)	Gender (Female/Male)	Participant Unique visit Group
3. Visual attention	Visit duration (ms)	3.1 Semantic similarity percentage (%) 3.2 Proportion of artefacts remembered (%) Number of artefacts used in the final build (N)	Gender (Female/Male)	Participant Unique visit Group

Table 4. Variable selection for the models.

4.3.2 Transformation of the variables

The linear mixed-effects model expects normally distributed variables, and therefore the shape of the distribution of the variables were measured with skewness and kurtosis. Skewness measures the asymmetry of the distribution, whereas kurtosis measures the tailedness of the distribution's central peak in relation to a normal distribution (Cain et al., 2017). All response variables were standardised using z-scores by subtracting the mean and dividing by the standard deviation for better interpretability. If the measured skewness of the variable was discovered to be over 1 and/or the kurtosis had a value over 3, a logarithmic transformation was applied before z-scoring (Cain et al., 2017).

4.3.3 Intercept-only models and random effects structure

To establish a baseline for predicting JVA and initiating JVA, logistic mixed-effects regression models were first fitted using only the dependent variable and random effects to account for variability in the occurrence of JVA and initiating JVA across different visits, participants, and groups. The intraclass correlation (ICC) value explains the proportion of variance in the model that can be attributed to differences between unique visits, participants and groups (McGraw & Wong, 1996).

The intercept-only model for JVA demonstrates that the ICC value for unique visits is 0.22%, indicating that only 0.22% of the variance in JVA can be attributed to differences between unique visits. The ICC value for participant is 3.15% and 15.01% for the group, suggesting that group differences contribute more substantially to the variability in JVA compared to the other factors. Based on these results, the unique visits were excluded from the random effects structure and a likelihood ratio test (LRT) (Peugh, 2010, p. 98) was conducted to compare the full random-effects model including unique visits as a random effect, to a reduced model with only groups and participants as random effects. The full random-effects model, which included visits, participants and groups as random effects, did not have a significantly better fit compared to the reduced model ($\chi^2(1) = 0.07$, $p = .788$).

The same is true for the intercept-only model for initiating JVA, where almost no variance can be attributed to unique visits (ICC ≈ 0.00%). The ICC value for the participant was 4.39%, and 7.7% for the group. Consequently, the random effect for unique visits was also removed from the model for predicting the initiation of JVA.

In a similar manner, a linear mixed-effects model was fitted to account for the variability in visit duration, with participants, groups and unique visits included as random effects. The ICC for the unique visit-level variance was 1.46%, and the ICC value for the participant-level variance was 4.98% and 3.41% for the group-level variance. However, the group was removed from the random effect structure because the model included the number of artifacts as an independent variable, which is a group-level variable. Additionally, the primary aim of the model was to predict the amount of visual attention of individual participants, and the low ICC value for the group-level variance indicated that only a small portion (3.41%) of the variance in visual attention could be attributed to differences between groups. Including the group-level random factor in the full model (see section 4.4.3) also resulted in convergence issues, specifically a singular fit. This made it necessary to simplify the random effects structure to achieve a stable and reliable model fit.

4.4 Statistical analysis and results

4.4.1 Joint visual attention and long-term memory

To test the hypothesis of whether JVA enhances narrative and/or visual long-term memory of the group's outcome, two binary GLMMs (narrative and visual) were fitted with a logistic link function using the lme4 package in R (Bates et al., 2015). The models were estimated using Maximum Likelihood Estimation (MLE) and BOBYQA optimizers. The results are presented in Table 5 and 6 with estimated coefficients (β), standard errors (SE), z-scores, p-values, and odds ratios with 95% confidence intervals (CI) for each predictor variable.

The narrative model's total explanatory power is substantial (conditional $R^2 = 0.27$), which means that the model explains 27% of the total variability in the narrative memory when considering both the fixed and the random effects. The part related to the fixed effects alone (marginal R^2) is 0.13. The visual model's total explanatory power is moderate (conditional $R^2 = 0.24$), and the fixed effects alone explain 12% of the variability in JVA (marginal $R^2 = 0.12$). The full validation of the models is presented in section 4.5.

Binary Logistic Regression Model Summary Joint Visual Attention and Narrative Long-Term Memory (Semantic Similarity Percentage)					
Predictor variables	**Coefficients** β	**SE**	**z-score**	**p-value**	**Odds Ratio (95% CI)**
Intercept (JVA)	-0.774	0.89	-0.87	.384	0.461 (0.081, 2.638)
Semantic similarity percentage	-0.195	0.109	-1.792	.073	0.823 (0.665, 1.019)
Visit duration	-0.196	0.064	-3.069	.002 (**)	0.822 (0.726, 0.932)
Number of fixations during a visit	0.81	0.064	12.594	< 2e-16 (***)	2.248 (1.982, 2.551)
Gender (female)	0.378	0.151	2.495	.013 (*)	1.459 (1.084, 1.964)
Group work duration (min)	-0.337	0.193	-1.743	.081	0.714 (0.489, 1.043)
Number of artefacts in the final build	-0.051	0.057	-0.895	.371	0.95 (0.85, 1.062)
Random effects	σ^2	**SD**			

Intercept (Participant)	0.021	0.145			
Intercept (Group)	0.64	0.8			
*** p < .001, ** p < .01, * p < .05					

Table 5. Joint visual attention and narrative long-term memory

The coefficient for narrative long-term memory is negative (β = -0.195, SE = 0.109, z = -1.792, p = .073), indicating that higher semantic similarity is associated with lower odds of JVA, although the effect is not statistically significant. The odds ratio of 0.823 indicates that for each unit increase in semantic similarity percentage, the odds of JVA decrease by approximately 17.7%. However, the 95% confidence interval of (0.665, 1.019) is very close to 1, which also means that this effect could be null.

The negative and statistically significant coefficient for visit duration (β = -0.196, SE = 0.064, z = -3.069, p < .01) indicates that longer visit durations are associated with a decrease in the odds of JVA. The odds of JVA decrease by approximately 17.8% for each one unit increase in visit duration (95%: CI 0.725, 0.932). Instead, the number of fixations is positive and highly significant coefficient (β =0.81, SE = 0.064, z = 12.594, p < .001), indicating that for each one unit increase in the number of fixations during a visit increases the odds of JVA by approximately 124.8% (95% CI: 1.982, 2.551). The positive and statistically significant coefficient for gender suggests that being a female increases the odds of JVA by approximately 45.9% compared to being male, assuming all other variables are held constant (95% CI: 1.084, 1.964). The duration of the group work task and the number of artefacts included in the final build did not have a significant effect on the odds of JVA. However, the negative coefficient for the number of artefacts in the final build suggest that as the number of artefacts increases, the odds of JVA decreases.

Binary Logistic Regression Model Summary Joint Visual Attention and Visual Long-Term Memory (Proportion of Artefacts Remembered)					
Predictor variables	**Coefficients *β***	**SE**	**z-score**	**p-value**	**Odds Ratio (95% CI)**
Intercept (JVA)	-0.947	0.764	-1.239	.215	0.388 (0.087, 1.735)

Proportion of artefacts remembered	0.05	0.101	0.499	.618	1.052 (0.863, 1.282)
Visit duration	-0.192	0.064	-3.016	.003 (**)	0.825 (0.728, 0.935)
Number of fixations during a visit	0.806	0.064	12.51	< 2e-16 (***)	2.239 (1.974, 2.541)
Gender (female)	0.364	0.204	1.788	.074	1.439 (0.966, 2.144)
Group work duration (min)	-0.362	0.17	-2.129	.033 (*)	0.7 (0.5, 0.972)
Number of artefacts in the final build	-0.034	0.049	-0.688	.491	0.967 (0.878, 1.065)
Random effects	**σ^2**	**SD**			
Intercept (Participant)	0.041	0.202			
Intercept (Group)	0.463	0.68			
*** p < .001, ** p < .01, * p < .05					

Table 6. Joint visual attention and visual long-term memory.

In the visual model, the coefficient for the proportion of artefacts remembered is positive (β = 0.05, SE = 0.101, z = 0.499), but it does not have a significant effect on the odds of JVA (p = .618). The positive coefficient and odds ratio slightly above 1 suggest a trend towards increased odds of JVA with higher proportion of artefacts remembered, but this trend is not statistically supported by the confidence interval (1.052, 95% CI: 0.863, 1.282). Similar to the narrative model, the visit duration shows a negative and statistically significant effect on the odds of JVA (β = -0.192, SE = 0.064, z = -3.016, p < .01), and the number of fixations has a positive and statistically significant effect on the odds of JVA (β = 0.806, SE = 0.065, z = 12.51, p < .001). In contrast to the narrative model, the visual model does not demonstrate a significant effect of gender on the odds of JVA. Instead, the analysis reveals a negative and statistically significant coefficient for the duration of group work tasks (β = -0.362, SE = 0.17, z = -2.129, p < .05), indicating that longer durations decrease the odds of JVA by approximately 30% (95% CI: 0.5, 0.972).

Based on these results, there is no significant evidence to support the enhancement of either narrative or visual long-term memory through JVA. Therefore, both null hypotheses of JVA enhancing narrative long-term memory (H1) and visual long-term memory (H2) cannot be rejected. However, both models indicate that an increase in the number of fixations during a visit increases the odds of JVA by approximately 123.9-124.8%. Moreover, both models show a statistically significant negative coefficient for the odds of JVA associated with longer visit durations (β = -0.196 and -0.192). The main differences and contradictory results in the models relate to the effects of gender and the duration of the group work task on the odds of JVA. The narrative model predicts that females are more likely to engage in JVA, whereas the visual model shows no significant effects on gender but predicts that longer group work task durations decrease the odds of JVA.

4.4.2 Initiating joint visual attention and long-term memory

To test the hypothesis whether initiating JVA is a better predictor of enhanced long-term memory of the group's outcome, two binary GLMMs (narrative and visual) were fitted using the lme4 package in R (Bates et al., 2015). The models were estimated using Maximum Likelihood Estimation (MLE) and BOBYQA optimizers. The narrative model's total explanatory power is moderate, with a conditional R^2 of 0.18. The part of the variance explained by the fixed effects alone, indicated by the marginal R^2, is 0.12. The visual models total explanatory power is moderate (conditional R^2 = 0.19), while the marginal R^2 is the same as in the narrative model (0.12). The results are presented in Table 7 and 8, and the full validation of the models is presented in section 4.5.

Binary Logistic Regression Model Summary
Initiating JVA and Narrative Long-Term Memory (Semantic Similarity Percentage)

Predictor variable	**Coefficients *β***	**SE**	**z-ratio**	**p-value**	**Odds Ratio (95% CI)**

Intercept (Initiating JVA)	-2.184	0.532	-4.102	4.10e-05 (***)	0.113 (0.04, 0.32)
Semantic similarity percentage	0.152	0.135	1.126	.26	1.164 (0.893, 1.516)
Visit duration	0.046	0.093	0.497	.619	1.047 (0.872, 1.258)
Number of fixations during a visit	0.597	0.091	6.589	4.42e-11 (***)	1.817 (1.521, 2.17)
Gender (female)	0.421	0.197	2.137	.033 (*)	1.523 (1.035, 2.241)
Group work duration (min)	-0.233	0.129	-1.809	.071	0.792 (0.62, 1.02)
Number of artefacts in the final build	-0.03	0.036	-0.842	.4	0.97 (0.905, 1.041)
Random effects	**σ^2**	**SD**			
Intercept (Participant)	0.035	0.186			

Intercept (Group)	0.2	0.448			
*** p < .001, ** p < .01, * p < .05					

Table 7. Initiating JVA and narrative long-term memory.

The intercept coefficient for the narrative model indicates that the log-odds of initiating JVA are significantly decreased by approximately 88.7% (95% CI: 0.04, 0.32). The coefficient for semantic similarity percentage suggests a positive association between initiating JVA, but it is statistically non-significant ($\beta = 0.152$, SE = 0.135, z = 1.126, p = .26). The odds ratio of 1.164 also indicates a slight increase in odds, but it is not significantly different from 1, suggesting no significant effect (1.164, 95% CI: 0.893, 1.516). Similarly to the model predicting JVA, the number of fixations shows a statistically significant positive effect also on the odds of initiating JVA ($\beta = 0.597$, SE = 0.091, z = 6.589, $p < .001$), with an odds ratio of 1.817. This means that for each one unit increase in the number of fixations, the odds of initiating JVA increase by approximately 81.7% (95% CI: 1.521, 2.17). Similarly to the JVA model on narrative memory, the coefficient estimate for female gender is positive and statistically significant ($\beta = 0.421$, SE = 0.197, z = 2.137, $p < .05$), suggesting that females have approximately 52.3% higher odds of initiating JVA compared to males (95% CI: 1.035, 2.241).

Binary Logistic Regression Model Summary Initiating JVA and Visual Long-Term Memory (Proportion of Artefacts Remembered)					
Predictor variable	**Coefficients *β***	**SE**	**z-ratio**	**p-value**	**Odds Ratio (95% CI)**
Intercept (initiating JVA)	-2.097	0.574	-3.65	.000 (***)	0.123 (0.04, 0.379)
Proportion of artefacts remembered	0.19	0.094	2.028	.043 (*)	1.21 (1.006, 1.454)

Visit duration	0.04	0.093	0.426	.67	1.041 (0.867, 1.25)
Number of fixations during a visit	0.605	0.091	6.641	3.11e-11 (***)	1.831 (1.532, 2.19)
Gender (female)	0.293	0.198	1.480	.139	1.34 (0.909, 1.974)
Group work duration (min)	-0.233	0.137	-1.706	.088	0.792 (0.606, 1.035)
Number of artefacts in the final build	-0.032	0.038	-0.847	.397	0.969 (0.9, 1.043)
Random effects	**σ^2**	**SD**			
Intercept (Participant)	0.026	0.161			
Intercept (Group)	0.252	0.502			
*** p < .001, ** p < .01, * p < .05					

Table 8. Initiating JVA and visual long-term memory.

In the visual memory model (Table 8), the positive coefficient for the proportion of artefacts remembered shows a statistically significant positive association with initiating JVA ($\beta = 0.19$, $SE = 0.094$, $z = 2.028$, $p < .05$). The odds ratio of 1.21 indicates that for each one-unit increase in the percentage of artefacts remembered, the odds of initiating JVA increase by approximately 21% (95% CI: 1.006, 1.454). Consistent with other models, the positive coefficient for the number of fixations during a visit indicates a strong and statistically significant positive association with the odds of initiating JVA ($\beta = 0.605$, SE = 0.091, $z = 6.641$, $p < .001$). However, visit duration, gender, duration of the group work task and the number of artefacts in the final build do not show significant effects on the odds of initiating JVA.

In light of these results, it can be concluded that initiating JVA does not enhance narrative long-term memory of the group's outcome, and the null hypothesis cannot be rejected (H3). However, initiating JVA does enhance the visual long-term memory of the group's outcome, allowing for the rejection of the null hypothesis (H4).

4.4.3 Visual attention and long-term memory

To analyse the sole effect of visual attention on visual and narrative long-term memory, two linear mixed-effects (LME) models were fitted with visit duration as a dependent variable (see Tables 9 and 10). The random effects included in the models were unique visits and participants, as justified in the section 4.3.3.

Linear Mixed-Effects Model Summary
Visual Attention and Narrative Long-Term Memory (Semantic Similarity Percentage)

Fixed effects	**Estimate**	**SE**	**t-value**	**p-value**
Intercept (Visual attention)	-0.202	0.192	-1.05	.312
Semantic similarity percentage	0.02	0.069	0.287	.779
Gender (female)	-0.353	0.17	-2.073	.057
Number of artefacts in the final build	0.026	0.016	1.672	.117

Random effects	**σ^2**	**SD**		
Intercept (Unique visit)	0.015	0.121		
Intercept (Participant)	0.069	0.263		
Observations 7808 AIC 21648 BIC 21697 logLik -10817 *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$				

Table 9. Visual attention and narrative long-term memory.

Linear Mixed-Effects Model Summary Visual Attention and Visual Long-Term Memory (Proportion of Artefacts Remembered)				
Fixed effects	**Estimate**	**SE**	**t-value**	**p-value**
Intercept (Visual attention)	-0.141	0.163	-0.864	.402
Proportion of artefacts remembered	0.121	0.066	1.825	.09
Gender (female)	-0.32	0.139	-2.314	.036 (*)

Number of artefacts in the final build	0.023	0.012	1.841	.087
Random effects	σ^2	**SD**		
Intercept (Unique visit)	0.015	0.121		
Intercept (Participant)	0.056	0.237		
Observations 7808 AIC 21644 BIC 21693 logLik -10815 *** p < 0.001, ** p < 0.01, * p < 0.05				

Table 10. Visual attention and visual long-term memory.

For the visual memory model, the proportion of remembered artefacts shows a positive but not statistically significant association with visit duration ($\beta = 0.121$, SE = 0.066, t = 1.825, p = 0.09). The same is true for the number of artefacts in the final build, with a coefficient of 0.023 (SE = 0.012, t = 1.841, p = 0.087), indicating a small positive but non-significant association with visit duration. While these effects are not statistically significant at the conventional 0.05 level, they are marginally significant ($p < 0.1$), suggesting potential trends that support the hypothesis of artefacts directing visual attention and the relationship between attention and long-term memory. Contrary to the JVA models, the female gender shows a significant negative coefficient (-0.32, SE = 0.139, t = -2.314, $p < 0.05$), indicating that being a female is associated with a decrease in the odds of longer visit duration. Adding the duration of the group work task as a predictor did not significantly improve the model fit at the conventional significance level ($p < 0.05$). Therefore, it was excluded from the model, which subsequently yielded a better BIC value of 21693, while the AIC value remained the same at 21644.

In the narrative memory model, the coefficient for semantic similarity percentage is 0.02 (SE = 0.069, t = 0.287) with a p-value of .779, indicating it is not statistically significant. Similarly to the visual model, the coefficient for female gender is -0.353 (SE = 0.17, t = -2.073), which is marginally significant at the p <

0.1 level (p = .057), supporting the finding that being female decreases the odds of visit duration. The number of artefacts in the final build is not statistically significant but suggests a potential positive trend (β = 0.026, SE = 0.016, t = 1.672, p = .117).

Based on these results, the null hypothesis that visual attention enhances narrative long-term memory (H1) cannot be rejected. While there is a marginally significant (p < 0.1) association observed between visual attention and visual long-term memory, suggesting some evidence against the null hypothesis (H2), this evidence is not strong enough to confidently reject it. Therefore, the null hypothesis that visual attention does not enhance visual long-term memory (H2) also cannot be rejected based on this study.

4.5 Validation of the models

To evaluate the goodness-of-fit and identify any potential issues with the fitted models, several diagnostic tests were performed using the DHARMa and pROC packages in R. The DHARMa package creates simulated residuals for non-parametric simulation-based diagnostics. By simulating the residuals, DHARMa allows checking the distributional assumptions of the residuals and identifying potential problems like overdispersion, zero-inflation, or outliers (Hartig, 2022). The Kolmogorov-Smirnov (KS) test for residual uniformity, a nonparametric dispersion test, and an outlier test were performed for model diagnostics. The KS test evaluates whether the scaled residuals from the model follow a uniform distribution, as expected under the null hypothesis of no model misfit (Hartig, 2022). The KS test result indicates the maximum difference between the empirical cumulative distribution function (ECDF) of the residuals and the theoretical cumulative distribution function (CDF) of the reference distribution (Ghasemi & Zahediasl, 2012). The nonparametric dispersion test detects overdispersion in the residuals, where values around 1 suggest no overdispersion. The outlier test identifies whether there are more simulation outliers in the residuals than expected (Hartig, 2022).

Finally, to evaluate the discrimination ability of the binary logistic GLMM models and the performance of the binary classifiers, areas under the Receiver Operating Characteristic (ROC) curve (AUC) were calculated. The AUC value ranges from 0.5 to 1.0, where 0.5 indicates no discrimination (equivalent to random guessing), while a value of 1.0 indicates perfect discrimination (Lu & Scott, 2023). Additionally, the BIC (Bayesian Information Criterion) and AIC (Akaike Information Criterion), although primarily used for model comparison, are reported here for transparency, where lower values generally indicate a better model fit. The results are presented in Table 11.

	GLMM	LME

Model/ Test	JVA and narrative long-term-memory	JVA and visual long-term-memory	Initiating JVA and narrative long-term-memory	Initiating JVA and visual long-term memory	Attention and narrative long-term memory	Attention and visual long-term memory
KS	0.027 (***)	0.024 (***)	0.015 (.059)	0.012 (.226)	0.042 (***)	0.047 (***)
Dispersion	1.094 (.456)	1.066 (.592)	1.085 (.456)	1.062 (.616)	0.983 (0.52)	0.989 (.72)
Outliers	67/7808 (.524)	61/7808 (.949)	76/7808 (.085)	71/7808 (.252)	41/7808 (***)	37/7808 (***)
AUC	0.71	0.712	0.717	0.717	-	-
BIC	8491	8494.1	5196.6	5193.7	21724.5	21721.7
AIC	8429.3	8431.5	5133.9	5131	21668.7	21666
*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$						

Table 11. Validation of the models.

Although the scaled residuals (KS test) for the JVA and visual attention models do not follow a uniform distribution, the dispersion test indicates no significant underdispersion or overdispersion of the residuals. Additionally, given the large sample size (7808 data points), it is expected that the signal-to-noise ratio is higher, which can make it easier to detect deviations in the residuals (Hartig, 2022). The statistically significant but proportionally small outliers detected in the linear mixed-effects models for predicting visual attention (0.52% and 0.47%) could be attributed to measurement errors in the eye-tracking data, particularly in detecting unique visits. All validation results were also visualised to detect any clear patterns in the residuals as suggested by Hartig (2022), and the visual inspection did not reveal any significant concerns. The example inspection of the visualisations is presented in Appendix 2.

5 DISCUSSION

Based on the findings of this study, JVA does not enhance either narrative or visual long-term memory of the group's outcome. Additionally, the coefficient for narrative memory is negative, suggesting that increased JVA decreases narrative long-term memory.

The dynamic analysis of JVA suggests that initiating JVA enhances visual long-term memory of the group's outcome, but this effect does not extend to narrative long-term memory. Previous research has found that initiating JVA facilitates episodic memory retrieval (Gomez et al., 2009), hypothetically due to the role of egocentric spatial information processing (Kim & Mundy, 2012). However, the theory does not directly explain why visual long-term memory is enhanced more than narrative long-term memory, as the recollection of narrative structure is considered a higher-level aspect of episodic memory (Cohn-Sheehy et al., 2021).

Interestingly, the findings of this study suggest that the effect of JVA on narrative long-term memory is negative. The negative effect could be explained by factors related to the group's outcome, such as its innovativeness or coherence. According to Neisser's (1967) schema theory, existing mental schemas are important for encoding narrative information (Masís-Obando et al., 2022). The application of this theory suggests that if the group's outcome does not properly align with existing cognitive schemas, it will not get stored in long-term memory. Hence, although tactile artefacts have been presented as co-created knowledge objects, analogous to shared cognitive schemata, their contribution to knowledge integration might not be as straightforward as presented (see e.g., Teh et al., 2021). Therefore, in case the group's outcome was indeed perceived as innovative, it could explain the negative effect of JVA on narrative long-term memory if the outcome did not align with existing mental schemas.

To explain the possible reasons for the differences in the effects of JVA and initiating JVA, it has been studied that the volitional control related to initiating JVA recruits neural reward circuits in ventral striatum (Schilbach et al., 2020). Hence, initiating JVA may enhance memory of the attentional target by releasing dopamine in hippocampus, which is involved in the process of memory consolidation (Adcock et al., 2016, Shohamy & Adcock, 2010). This neurophysiological process could also explain why visual long-term memory is enhanced when initiating JVA compared to narrative long-term memory, as visual attention is primarily directed towards artefacts, aiding in the encoding and consolidation of visual information.

The lower memory performance in narrative long-term memory might also be explained by differences in brain networks involved in the encoding and retrieval of narratives (Masís-Obando et al., 2022). Narratives typically get consolidated with specific attachment points, which in this study are concretely presented by the tactile artefacts. However, if the retrieval process of a narrative relies on a sequential memory cueing strategy—where participants first remember the events of the story and then its related details (Masís-Obando et al., 2022)—the artefacts intended to embody the story might hinder the retrieval of the narrative. This effect could be amplified by the experimental design of this study, in which participants were first asked to remember the artefacts and their assigned meanings, and only afterward, the final story (see Appendix 1).

The results from the other attentional metrics used in this study support the dynamic and complex nature of JVA. For example, the statistically significant positive coefficient for the number of fixations in predicting both JVA and initiating JVA suggests that maintaining JVA requires adapting to social cues, which can be hindered by fewer fixations. This hypothesis is further supported by the finding that longer visit durations appear to negatively impact the odds of JVA. However, visit duration alone as a metric cannot reliably predict the probability of JVA because it does not capture the dynamics of eye movements related to JVA. Instead, it serves as a reasonable indicator of sustained attention (Styles, 2005), which is necessary for the consolidation of information into long-term memory. Additionally, visit and fixation durations must be interpreted with caution, as they can also indicate mind-wandering (deBettencourt et al., 2021). In summary, while longer visit durations positively affect the odds of initiating JVA, the positive coefficient also for the number of fixations highlights the importance of detecting fixation rates over visit and fixation durations.

The main differences on predicting the odds of JVA found in this study pertain to the effects of gender and the duration of the group work task. The narrative long-term memory model of JVA predicts that females are more likely to engage in JVA, whereas the visual long-term memory model shows no significant effect on gender but predicts that longer group work task durations decrease the odds of JVA. The negative coefficient for the duration of the group work task may be attributed to participants becoming fatigued during the workshop and suggests that a time-based analysis might be needed to understand the fluctuations in JVA over the course of the task. The coefficient of being female increased the odds of both JVA (odds = 45.9%, $p < .05$, odds = 43.9%, $p = .074$) and initiating JVA (odds = 52.3%, $p < .05$, odds = 34%, $p = .139$). This is consistent with previous research on JVA, which has found that the reflexive shift of attention, tested with the original gaze cueing paradigm, is stronger in females than in males (Bayliss et al., 2005). In addition, while this observation must be interpreted with caution due to the low number of female subjects, it does provide further validation for the developed JVA detection algorithm and the accuracy of the eye-tracking data.

The negative, although statistically non-significant, coefficient for the number of artefacts in the final build suggests that as the number of artefacts increases, the odds of JVA decreases. This could possibly be due to participants' attention becoming more divided. However, determining the optimal number of artefacts for enhancing joint visual attention, collaboration quality, and long-term memory is challenging and may not be worth further deliberation, as the results suggest no statistically significant effect for the number of artefacts on the odds of JVA.

5.1 Limitations

To address the thought-provoking results related to the effect of JVA on narrative long-term memory, it might be necessary to critically examine the methodological approach of comparing the semantic similarity between the long-term memory interview answers and the final recording of the group’s outcome. Initially, instead of conducting online follow-up interviews, follow-up long-term questionnaires were sent to the participants from the first two groups as the conducted pilot studies failed to indicate their unsuitability for this study. Only after receiving responses to the long-term memory questionnaires from the first two groups, the questionnaire was replaced by an online long-term memory interview. The first two groups were also interviewed online, but as a part of the follow-up questionnaire, they had already seen the picture of the final build.

In addition, although all groups were instructed similarly to create the final video, where only one person from each group described the meaning of the final build, there may be variation in how thoroughly the final build was described. This variation might also favour the recall of the final build's meaning by those who described it, but this potential effect was not assessed in this study. Hence, although the direct comparison of the cosine-similarity between the transcribed final video and the long-term memory interview answers is probably the most objective method for analysing the data, the experimental design sets its own limitations to reliably study narrative long-term memory. Moreover, assuming a positive correlation between narrative and visual long-term memory and JVA ignores the time in between the workshop and the follow-up interview, where the information either gets or does not get consolidated into long-term memory. For example, a few bad nights of sleep or increased stress after the workshop can hypothetically have a much stronger effect on long-term memory than the amount of JVA or visual attention during the workshop.

Something that could have affected the long-term memory of the group's outcome, but not accessed in this study, is tied to the experienced level of innovativeness and coherency of the outcome. As the groups were recruited from within companies, the participants presumably knew each other—and the company’s practices—well. Therefore, it is possible that some of the groups were less innovative in developing a refined common understanding on how to enhance customer engagement and instead, settled for describing the current processes. This might partially explain the implicit observations from the long-term memory interviews that although the tactile artefacts were not always remembered, the final story was. It is known that people within organisations often adhere closely to established organisational processes for organisation stability, which can also act as a barrier to innovation. Therefore, although it can be theoretically argued that remembering the outcome is associated with the power of storytelling, visual information and joint visual attention, it would have been beneficial to have the participants rate the group's outcome in terms of its innovativeness to rule out the familiarity effect. However, the results suggest that performance for narrative long-term memory was not as strong as for visual long-term memory, which could indirectly indicate that the outcome was innovative. Another explanation relates to the recall theory of narratives, suggesting that the sequential memory cueing strategy may have been disrupted when participants were asked to recall the artefacts themselves, rather than the 'events' of the strategy (Masís-Obando et al., 2022). In fact, another important factor that would have been beneficial to add in the questionnaire, is related to the coherency of the outcome as research shows that recall is enhanced when separate events are integrated into a coherent narrative (Cohn-Sheehy et al., 2021). Although the iterative working method is designed to ensure that everyone's opinion is heard, it does not guarantee that all participants will rate the outcome as equally coherent, which could affect their ability to recall the outcome. In hindsight, adding a subjective rating of innovativeness and coherency of the group's outcome would have allowed a more reliable evaluation of the effect of JVA on long-term memory and helped rule out other, potentially influencing, factors.

It should be also highlighted that (joint) visual attention was analysed solely at the space-level with the help of separate AOIs. The object-level approach could potentially provide more detailed data on visual attention allocation towards specific artefacts, allowing a more reliable analysis of whether the specific attended artefacts were also remembered. However, in this study, the object-level analysis was not considered as crucial and applicable because the artefacts were also verbally referenced, activating both visuospatial and verbal working memories along with narrative structures. Consequently, the chosen space- and context-level analysis for (joint) visual attention was aligned with the long-term memory analysis by taking into account only those visually remembered artefacts whose assigned meanings were also remembered.

Finally, it must be noted that the study did not involve control groups using more conventional collaboration methods and aids, such as textual communication, so no explicit conclusions can be drawn about visual communication aids.

5.2 Implications for future research

Future research should continue to study JVA in real-life scenarios to further investigate the dynamic nature of JVA and its indicators. Further validation and contributions are especially needed for detecting JVA events in natural interactions involving more than two participants. In this study, a previously validated time threshold of ±2 seconds (Richardsson & Dale, 2005) was applied to individual fixations to detect JVA, yielding a reasonable average duration of JVA events.[3] However, no additional tests were developed or performed to find the most reliable time windows for detecting JVA events.

The design language of the tactile artefacts was not assessed in this study, but future research could focus on studying whether certain features of the artefacts direct more visual attention. Hypothetically, more visually rich and complex artefacts might attract more visual attention; however, they could also impede the metaphorical thinking that the artefacts are intended to encourage. In addition, more research is needed to find a balance between providing enough artefacts to engage participants in JVA and avoiding an excess that could lead to fragmented visual attention and, consequently, diminished long-term memory. Examining these dynamics in various contexts and with larger, more diverse samples could provide deeper insights and help refine the application of JVA and artefacts in enhancing collaboration and long-term memory. Future studies focusing on JVA and long-term memory should also consider a more controlled experimental design, where the levels of stress and fatigue are assessed both during and after the workshop.

To explain the negative coefficient for narrative long-term memory, a more detailed analysis of the eye-tracking data might be needed to compare instances of listeners looking at the artefacts versus looking at the speakers while the speakers make references to the artefacts. This highlights the role of selective attention in interaction, as visual attention to the speaker rather than the referenced object is crucial for comprehension and encoding, allowing listeners to capture non-verbal cues such as facial expressions. The effect of JVA on enhancing memory has been explained through the depth of information processing, where both one's own visual attention and others' visual attention are processed (Kim & Mundy, 2012). However, applying this theory in natural face-to-face interaction studies needs some further refinement, as it does not account for instances when people are looking at each other.

An unconventional suggestion for the experimental design, focusing on studying narrative long-term memory in a collaborative context, is to incorporate dynamic visual content or virtually animated artefacts. For instance, applying

---

[3] The mean JVA event duration on the joint building board was 10.3 seconds (SD = 13.4). The maximum physical distance between the visits during an event was 41 cm and the maximum JVA event duration was 110.6 seconds. The mean of the maximum distance between visits during a JVA event on the joint building board was 9.3 cm (SD = 5.1).

neuropsychological findings on narrative encoding and retrieval (Masís-Obando et al., 2022) suggests that if participants had moved the artefacts more frequently during the workshop—particularly when describing the meaning of the final build—it could enhance narrative retrieval by encoding those actions as 'events'. An implicit finding from the long-term memory interviews supports this hypothesis on motion visuals, as some participants attempted to recall the outcome by remembering the actions when someone placed a certain artefact on the table.

5.3 Implications for practise

The results of this study indicate that JVA enhances long-term visual memory of the group's outcome compared to visual attention alone. It can be fairly confidently stated that tactile artefacts are effective at directing JVA, provided the number of artefacts is carefully assessed. The direction of the results indicates that increasing the number of tactile artefacts decreases JVA; however, no direct suggestions can be made for the optimal number of artefacts to enhance JVA, collaboration quality and long-term memory.

Since the results do not support the hypothesis that JVA enhances the narrative long-term memory of the group's outcome, adding visual communication aids in strategy work must be carefully evaluated before its implementation. Some implicit observations from the long-term memory interviews indicate that remembering the individual artefacts and their assigned meanings was not necessary for successfully recalling the meaning of the final build. Conversely, successfully recalling the artefacts did not ensure the successful recall of the final build's meaning. In practice, if the visual information provided for collaboration is crucial to remember, directing JVA towards it is recommended. However, if the visual information is only meant to support the collaborative strategy work, the results suggest that the amount of visual attention or JVA is not essential for successfully retrieving the strategy. Although the optimal type of visual information to support strategy work remains unresolved and was not covered in this study, tactile artefacts as adaptable and movable visuals appear to be a promising solution.

6 CONCLUSION

The aim of this study was to investigate if and how (joint) visual attention in a collaborative face-to-face interaction enhances long-term memory of the group's outcome. To accurately study the occurrences of (joint) visual attention and its effect on long-term memory, the experimental design was built upon tactile artefacts and eye-tracking methodology. Directing (joint) visual attention towards tactile artefacts and using them to build a common understanding was assumed to facilitate information encoding by activating visuospatial working memory, and creating a set anchor points for retrieving the outcome in a narrative format. Previous research has found that JVA during a collaborative context supports learning (Schneider & Pea, 2013) and enhances collaboration (Schneider et al., 2018, Schneider & Bryant, 2022). However, the effect of JVA on long-term memory has received less attention, despite subsequent evidence that JVA enhances visual working memory (Gregory & Jackson, 2019, Gregory & Jackson, 2017, Gregory & Kessler, 2022). The study and its methodology were seen as valuable contributions to the research field of JVA, which is still trying to validate the laboratory findings of JVA effects in real-life scenarios (Caruana et al., 2017). Additionally, the practical motivation of this study was to explore the role of visual information in improving strategy work and collaboration quality.

A total of eight groups participated in the study, which employed a nested design where the groups of three to four people took part in a facilitated workshop focused on enhancing customer engagement. All groups collaborated using tactile artifacts and were encouraged to incorporate storytelling practices in describing the jointly built strategy. The results of this study do not show a clear effect of JVA on either improving the visual long-term memory of the tactile artefacts themselves or on the narrative long-term memory related to the meaning of selected artefacts and to the commonly built strategy on enhancing customer engagement. However, the dynamic analysis of JVA by detecting the initiators of JVA revealed that visual long-term memory increases the odds of initiating JVA. This finding is consistent with previous research, which has found that initiating JVA enhances recognition memory of pictures compared to responding to JVA (Kim & Mundy, 2012). The results of this study indicate that initiating JVA enhances visual long-term memory even in the free recall condition, as the data was analysed exclusively from this condition and still showed an effect.

An additional key insight of this study is related to the number of fixations, which increase the odds of JVA and initiating JVA. In contrast, longer visit durations decrease the odds of both. This finding could be explained by the fact that successful social communication demands continuous reading and adaptation to social cues, such as eye gaze, which is harder to accomplish with longer visit durations and fewer number of fixations.

Another noteworthy observation is the effect of gender, as females seem to have higher odds of both JVA and initiating JVA. This conforms to existing research indicating that females have an advantage in judging nonverbal cues (Hall, 1978), which is also a prerequisite for JVA (Bayliss et al., 2005). However, this finding must be interpreted with caution, considering the relatively small sample size of this result ($N = 7$).

Therefore, although the early eye-mind hypothesis states that longer average fixation duration equals deeper processing, this relationship does not seem to hold true with longer visit durations and long-term memory within a collaborative context. Additionally, although the dual coding theory suggests that tactile artefacts and storytelling could enhance long-term memory independently of their use in JVA, the results of this study indicate that JVA provides a stronger enhancement of visual long-term memory compared to visual attention alone.

REFERENCES

Abdu, R., & Schwarz, B. (2020). Split up, but stay together: Collaboration and cooperation in mathematical problem solving. *Instructional Science*, 48(3), 313–336.

Adcock, R. A., Thangavel, A., Whitfield-Gabrieli, S., Knutson, B., & Gabrieli, J. D. (2006). Reward-motivated learning: Mesolimbic activation precedes memory formation. *Neuron*, 50(3), 507-517.

Allen, J. , Thiese, M. , Eden, E. & Knowles, S. (2022). Why Am I So Exhausted?. *Journal of Occupational and Environmental Medicine*, 64 (12), 1053-1058.

Awh, E., Vogel, E. K., & Oh, S. H. (2006). Interactions between attention and working memory. *Neuroscience*, 139(1), 201–208.

Baddeley, A. D., & Warrington, E. K. (1973). Memory coding and amnesia. *Neuropsychologia*, 11(2), 159–165.

Baldauf, D., & Desimone, R. (2014). Neural mechanisms of object-based attention. *Science* 344, 424–427.

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).

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1–48.

Bayliss, A. P., di Pellegrino, G., & Tipper, S. P. (2005). Sex differences in eye gaze and symbolic cueing of attention. *The Quarterly Journal of Experimental Psychology. A, Human experimental psychology*, 58(4), 631–650.

Bayliss, A. P., Frischen, A., Fenske, M. J., & Tipper, S. P. (2007). Affective evaluations of objects are influenced by observed gaze direction and emotional expression. *Cognition*, 104, 644–653.

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences of the United States of America*, 105(38), 14325–14329.

Brennan, S. E., Chen, X., Dickinson, C. A., Neider, M. B., & Zelinsky, G. J. (2008). Coordinating cognition: the costs and benefits of shared gaze during collaborative search. *Cognition*, 106(3), 1465–1477.

Brône, G., & Oben, B. (Eds.). (2018). *Eye-tracking in interaction: Studies on the role of eye gaze in dialogue.* John Benjamins Publishing Company.

Cain, M. K., Zhang, Z., & Yuan, K. H. (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior research methods*, 49(5), 1716–1735.

Cavanagh, P., Caplovitz, G.P., Lytchenko, T.K. et al. (2023). The Architecture of Object-Based Attention. *Psychon Bull Rev* 30, 1643–1667.

Carrasco, M. (2011). Visual attention: the past 25 years. *Vision research*, 51(13), 1484–1525.

Caruana, N., McArthur, G., Woolgar, A., & Brock, J. (2017). Simulating social interactions for the experimental investigation of joint attention. *Neuroscience and biobehavioral reviews*, 74(Pt A), 115–125.

Chen, Z. (2012). Object-based attention: a tutorial review. *Attention, perception & psychophysics*, 74(5), 784–802.

Chuang, C.-H., & Hsu, H.-C. (2023). Pseudo-mutual gazing enhances interbrain synchrony during remote joint attention tasking. *Brain and Behavior*, 13, e3181.

Clark, A. M., Intoy, J., Rucci, M., & Poletti, M. (2022). Eye drift during fixation predicts visual acuity. *Proceedings of the National Academy of Sciences of the United States of America*, 119(49), e2200256119.

Cohen, R. A. (1993). *The Neuropsychology of Attention*. Plenum Press, New York and London.

Cohn-Sheehy, B. I., Delarazan, A. I., Crivelli-Decker, J. E., Reagh, Z. M., Mundada, N. S., Yonelinas, A. P., Zacks, J. M., & Ranganath, C. (2022). Narratives bridge the divide between distant events in episodic memory. *Memory & cognition*, 50(3), 478–494.

Cole, G.G., Smith, D.T. & Atkinson, M.A (2015). Mental state attribution and the gaze cueing effect. *Atten Percept Psychophys,* 77, 1105–1115.

Cowan, N., Bao, C., Bishop-Chrzanowski, B.M., Costa, A. N., Greene N. R., Guitard, D., Li., C., Musich, M. L. & Ünal, Z. E. (2023). The Relation Between Attention and Memory. *Annu. Rev. Psychol*. 75:4.1–4.32

Dodd, M. D., Weiss, N., McDonnell, G. P., Sarwal, A., & Kingstone, A. (2012). Gaze cues influence memory…but not for long. *Acta psychologica*, 141(2), 270–275.

Doherty-Sneddon, G., Bruce, V., Bonner, L., Longbotham, S., and Doyle, C. (2002). Development of gaze aversion as disengagement from visual information. *Dev. Psychol*. 38:438.

Doherty-Sneddon, G., & Phelps, F. G. (2005). Gaze aversion: a response to cognitive or social difficulty? *Memory Cognit*. 33, 727–733.

deBettencourt, M. T., Williams, S. D., Vogel, E. K., & Awh, E. (2021). Sustained Attention and Spatial Attention Distinctly Influence Long-term Memory Encoding. *Journal of cognitive neuroscience*, 33(10), 2132–2148.

Duncan J. (1984). Selective attention and the organization of visual information. Journal of experimental psychology. *General*, 113(4), 501–517.

Entwistle, N., & Marton, F. (1994). Knowledge objects: Understandings constituted through intensive academic study. *British Journal of Educational Psychology*, 64(1), 161–178.

Fındık-Coşkunçay, D. & Çakır, M.P. (2022). An investigation of the relationship between joint visual attention and product quality in collaborative business process modeling: a dual eye-tracking study. *Softw Syst Model* 21, 2429–2460.

Fiol, C. M., & Huff, A. S. (1992). Maps for managers: Where are we? Where do we go from here? *Journal of Management Studies*, 29(3): 267–2.

Friesen, C.K. & Kingstone, A (1998). The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. *Psychonomic Bulletin & Review*, 5, 490–495.

Frintrop, S. (2006). 2 Background on Visual Attention. In *VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search. Lecture Notes in Computer Science*, vol 3899. Springer, Berlin, Heidelberg.

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.

Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., & Tylén, K. (2012). Coming to Terms: Quantifying the Benefits of Linguistic Coordination. *Psychological Science*, 23(8), 931-939.

Gallup, A. C., Hale, J. J., Sumpter, D. J., Garnier, S., Kacelnik, A., Krebs, J. R., & Couzin, I. D. (2012). Visual attention and the acquisition of information in human crowds. *Proceedings of the National Academy of Sciences of the United States of America*, 109(19), 7245–7250.

Garrod, S. , & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences*, 8(1), 8–11.

Gergle, D., Kraut, R. E., & Fussell, S. R. (2013). Using visual information for grounding and awareness in collaborative tasks. *Human – Computer Interaction*, 28(1), 1–39.

Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: a guide for non-statisticians. *International journal of endocrinology and metabolism*, 10(2), 486–489.

Gilboa, A. (2004). Autobiographical and episodic memory and the same? Evidence from prefrontal activation in neuroimaging studies. *Neuropsychologia*, 42, 1336–1349.

Glenberg, A. M., Schroeder, J. L., and Robertson, D. A. (1998). Averting the gaze disengages the environment and facilitates remembering. *Memory Cogn*. 26, 651–658.

Gomez, A., Rousset, S., & Baciu, M. (2009). Egocentric-updating during navigation facilitates episodic memory retrieval. *Acta Psychol*. 132, 221–227.

Gregory, S. E. A., & Jackson, M. C. (2017). Joint attention enhances visual working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(2), 237–249.

Gregory, S. E. A., & Jackson, M. C. (2019). Barriers block the effect of joint attention on working memory: Perspective taking matters. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 45(5), 795–806.

Gregory, S. E., & Kessler, K. (2022). Investigating age differences in the influence of joint attention on working memory. *Psychol. Aging*. 37, 731–741.

Großekathöfer, J. D., Suchotzki, K. & Gamer, M. (2020). Gaze cueing in naturalistic scenes under top-down modulation – Effects on gaze behaviour and memory performance. *Visual Cognition*, 28:2, 135-147.

Gruber, O., & von Cramon, D. Y. (2003). The functional neuroanatomy of human working memory revisited. Evidence from 3-T fMRI studies using classical domain-specific interference tasks. *NeuroImage*, 19(3), 797–809.

Hall, J. A. (1978). Gender effects in decoding nonverbal cues. *Psychological Bulletin*, 85(4), 845–857.

Hartig, F. (2022, September 8). DHARMa: Residual diagnostics for hierarchical (multi-level/mixed) regression models. Theoretical Ecology, University of Regensburg. Retrieved July 18, 2024, from https://cran.r-project.org/web/packages/DHARMa/vignettes/DHARMa.html#binomial-data

Hessels, R. S., Teunisse, M. K, Niehorster, D. C., Nyström, M., Benjamins, J. S., Senju, A. & Hooge, I. T. C. (2023). Task-related gaze behaviour in face-to-face dyadic collaboration: Toward an interactive theory?. *Visual Cognition*, 31:4, 291-313.

Hirschstein, Z., & Aly, M. (2023). Long-term memory and working memory compete and cooperate to guide attention. *Attention, perception & psychophysics*, 85(5), 1517–1549.

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.

Hovhannisyan, M., Clarke, A., Geib, B. R., Cicchinelli, R., Monge, Z., Worth, T., Szymanski, A., Cabeza, R., & Davis, S. W. (2021). The visual and semantic features that predict object memory: Concept property norms for 1,000 object images. *Memory & Cognition*, 49(4), 712–731.

Itier, R. J., Villate, C., & Ryan, J. D. (2007). Eyes always attract attention but gaze orienting is task-dependent: evidence from eye movement monitoring. *Neuropsychologia,* 45(5), 1019–1028.

Jellema, T., Baker, C. I., Wicker, B., & Perrett, D. I. (2000). Neural representation for the perception of the intentionality of actions. *Brain and cognition*, 44(2), 280–302.

Jermann, P. & Nüssli, M-A. (2012). Effects of sharing text selections on gaze cross-recurrence and interaction quality in a pair programming task. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work (CSCW '12).* Association for Computing Machinery, New York, NY, USA, 1125–1134.

Jermann, P., Mullins, D., Nüssli, M.-A., & Dillenbourg, P. (2011). Collaborative gaze footprints: Correlates of interaction quality. *CSCL2011 Conference Proceedings*. Volume I - Long Papers, 184–191.

Kahneman, D. (1973). *Attention and effort*. PRENTICE-HALL, Inc., Englewood Cliffs, New Jersey.

Kapetanios, E., Tatar, D., & Sacarea, C. (2013). *Natural language processing: Semantic aspects*. Taylor & Francis Group.

Kim, K., & Mundy, P. (2012). Joint attention, social-cognition, and recognition memory in adults. *Frontiers in human neuroscience*, 6, 172.

Kuhn, G., Pagano, A., Maani, S., & Bunce, D. (2015). Age-related decline in the reflexive component of overt gaze following. *Quarterly journal of experimental psychology*, 68(6), 1073–1081.

Kuhn, G., Teszka, R., Tenaw, N., & Kingstone, A. (2016). Don't be fooled! Attentional responses to social cues in a face-to-face and video magic trick reveals greater top-down control for overt than covert attention. *Cognition*, 146, 136–142.

Krall, S. C., Rottschy, C., Oberwelland, E., Bzdok, D., Fox, P. T., Eickhoff, S. B., Fink, G. R., & Konrad, K. (2015). The role of the right temporoparietal junction in attention and social interaction as revealed by ALE meta-analysis. *Brain structure & function*, 220(2), 587–604.

Krekelberg, B. (2011). Microsaccades. *Current Biology*, 21(11).

Kristjánsson, Á. & Draschkow, D. (2021). Keeping it real: Looking beyond capacity limits in visual cognition. *Atten Percept Psychophys* 83, 1375–1390.

Lankow, J., Ritchie, J., & Crooks, R. (2012). *Infographics: The power of visual storytelling*. John Wiley & Sons, Incorporated.

León, C. (2016). An architecture of narrative memory. *Biologically Inspired Cognitive Architectures*, Volume 16, 19-33.

Lehmann-Willenbrock, N., Allen, J. A. & Kauffeld, S. (2013). A Sequential Analysis of Procedural Meeting Communication: How Teams Facilitate Their Meetings. *Journal of Applied Communication Research*, 41:4, 365-388.

Lu, Y., & Scott, M. (2023). Generalised Linear Mixture Model. Retrieved 18th July 2024 from https://bookdown.org/ks6017/GLM_bookdown3/

Lyapunov, S. I., Shoshina, I. I. & Lyapunov, I. S. (2022). Tremor Eye Movements as an Objective Marker of Driver's Fatigue. *Human Physiology*, 48, 71-77.

Manera, V., Elena, M. R., Bayliss, A. P., & Becchio, C. (2014). When seeing is more than looking: Intentional gaze modulates object desirability. *Emotion*, 14(4), 824–832.

Masís-Obando, R., Norman, K. A., & Baldassano, C. (2022). Schema representations in distinct brain networks support narrative memory during encoding and retrieval. *eLife*, *11*, e70445.

McCarley, J. S., Leggett, N., & Enright, A. (2021). Shared Gaze Fails to Improve Team Visual Monitoring. *Human Factors*, 63(4), 696-705.

McCusker, S., & Swan, J. (2018). The Use of Metaphors With LEGO® SERIOUS PLAY® For Harmony and Innovation. *International Journal of Management and Applied Research*, 5(4), 174-192.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1), 30-46, 30-46.

McKay, K. T., Grainger, S. A., Coundouris, S. P., Skorich, D. P., Phillips, L. H., & Henry, J. D. (2021). Visual attentional orienting by eye gaze: A meta-analytic review of the gaze-cueing effect. *Psychological Bulletin*, 147(12), 1269–1289.

Milgram, S., Bickman, L., & Berkowitz, L. (1969). Note on the drawing power of crowds of different size. *Journal of Personality and Social Psychology*, 13(2), 79–82.

Monsalves, M.J., Bangdiwala, A.S., Thabane, A. et al (2020). LEVEL (Logical Explanations & Visualizations of Estimates in Linear mixed models): recommendations for reporting multilevel data and analyses. *BMC Med Res Methodol*, 20(3).

Mundy, P., & Newell, L. (2007). Attention, Joint Attention, and Social Cognition. *Current directions in psychological science*, 16(5), 269–274.

Mundy P. (2018). A review of joint attention and social-cognitive brain systems in typical development and autism spectrum disorder. *The European Journal of Neuroscience*, *47*(6), 497–514.

Napier, R. & Sharp, E. (2019). *Not Just Another Meeting: Creative Strategies for Facilitation*. Association for Talent Development.

Neisser, U. (1967). *Cognitive Psychology*. PRENTICE-HALL, Inc., Englewood Cliffs, New Jersey.

Norris, D. (2017). Short-term memory and long-term memory are still different. *Psychological bulletin*, 143(9), 992–1009.

Nummenmaa, L. & Calder, A. J. (2009). Neural mechanism of social attention. *Trends in Cognitive Science*, 13(3), 135-143.

Olson, E. (2011). AprilTag: A robust and flexible visual fiducial system. *Proceedings - IEEE International Conference on Robotics and Automation.* 3400 - 3407.

Paivio, A. (2013). *Imagery and verbal processes*. Psychology Press.

Paxton, A., Roche, J. M., Ibarra, A., & Tanenhaus, M. K. (2021). Predictions of Miscommunication in Verbal Communication During Collaborative Joint Action. *Journal of speech, language, and hearing research: JSLHR*, 64(2), 613–627.

Perez-Osorio, J., Müller, H. J., Wiese, E. & Wykowska, A. (2015). Gaze Following Is Modulated by Expectations Regarding Others' Action Goals. *PLoS ONE 10(11).*

Peugh J. L. (2010). A practical guide to multilevel modeling. *Journal of school psychology*, 48(1), 85–112.

Rayner, K. (2009). The 35th Sir Frederick Bartlett Lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, 62(8), 1457-1506.

Redcay, E., Dodell-Feder, D., Pearrow, M. J., Mavros, P. L., Kleiner, M., Gabrieli, J. D., & Saxe, R. (2010). Live face-to-face interaction during fMRI: a new tool for social cognitive neuroscience. *NeuroImage*, 50(4), 1639–1647.

Reimers, N. & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Richardson, D.C. & Dale, R. (2005). Looking To Understand: The Coupling Between Speakers' and Listeners' Eye Movements and Its Relationship to Discourse Comprehension. *Cognitive Science*, 29, 1045-1060.

Richardson, D. C., Street, C. N. H., Tan, J. Y. M., Kirkham, N. Z., Hoover, M. A., & Ghane Cavanaugh, A. (2012). Joint perception: Gaze and social context. *Frontiers in Human Neuroscience*, 6, 194.

Roback, P., & Legler, J. (n.d.). Chapter 10: Logistic Regression. In *Beyond Multiple Linear Regression*. Retrieved July 18, 2024, from https://bookdown.org/roback/bookdown-BeyondMLR/ch-logreg.html

Rothman, D. (2021). *Transformers for Natural Language Processing : Build Innovative Deep Neural Network Architectures for NLP with Python, PyTorch, TensorFlow, BERT, RoBERTa, and More*. Packt Publishing.

Roos, J. and Victor, B. (2018). How It All Began: The Origins Of LEGO® Serious Play®, *International Journal of Management and Applied Research*, 5(4), 326-343.

Sasin, E. & Fougnie, D. (2021). The road to long-term memory: Top-down attention is more effective than bottom-up attention for forming long-term memories. *Psychon Bull Rev* 28, 937–945.

Saxe, R., Xiao, D. K., Kovacs, G., Perrett, D. I., & Kanwisher, N. (2004). A region of right posterior superior temporal sulcus responds to observed intentional actions. *Neuropsychologia*, 42(11), 1435–1446.

Schneider, B. & Bryant, T. (2024). Using Mobile Dual Eye-Tracking to Capture Cycles of Collaboration and Cooperation in Co-located Dyads, *Cognition and Instruction*, 42(1), 26-55.

Schneider, B., & Pea, R. (2013). Real-time mutual gaze perception enhances collaborative learning and collaboration quality. *International Journal of Computer-Supported Collaborative Learning*, 8(4), 375–397.

Schneider, B., Sharma, K., Cuendet, S. et al. (2018). Leveraging mobile eye-trackers to capture joint visual attention in co-located collaborative learning groups. *International Journal of Computer-Supported Collaborative Learning,* 13, 241–261.

Schilbach, L., Wilms, M., Eickhoff, S., Romanzetti, S., Tepest, R., Bente, G., Shah, N. J., Fink, G. R., & Vogeley, K. (2010). Minds made for sharing: initiating joint attention recruits reward-related neurocircuitry. *J. Cogn. Neurosci,* 22, 2702–2715.

Schulz, K.-P., Geithner, S., Woelfel, C. & Krzywinski, J. (2015). Toolkit-based modelling and serious play as means to foster creativity in innovation processes. *Creativity and Innovation Management*, 24:2, 323-340.

Serious Play Community. (n.d.). LEGO® SERIOUS PLAY® methodology. Serious Play Community. Retrieved July 24, 2024, from https://seriousplay.community/usa/methodology/

Shohamy, D., & Adcock, R. A. (2010). Dopamine and adaptive memory. *Trends in Cognitive Sciences*, 14(10), 464-472.

Siirtola, H., Špakov, O., Istance, H. & Räihä, K-J. (2019). Shared Gaze in Collaborative Visual Search. *International Journal of Human–Computer Interaction*, 35:18, 1693-1705.

Siposova, B., & Carpenter, M. (2019). A new look at joint attention and common knowledge. *Cognition*, 189, 260–274.

Slessor G., Phillips L. H., & Bull R. (2008). Age-related declines in basic social perception: Evidence from tasks assessing eye-gaze processing. *Psychology and Aging*, 23(4), 812–822.

Soto, D. & Blanco, M. J. (2004). Spatial attention and object-based attention: A comparison within a single task. *Vision Research*, 44(1), 69-81.

Star, S. L., & Griesemer, J. R. (1989). Institutional Ecology, "Translations" and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science*, 19(3), 387–420.

Styles, E. (2005). *Attention, perception and memory: An integrated introduction.* Taylor & Francis Group.

Sun, Z., Yu, W., Zhou, J. et al (2017). Perceiving crowd attention: Gaze following in human crowds with conflicting cues. *Atten Percept Psychophys*, 79, 1039–1049.

Tatler, B. W., Gilchrist, I. D., & Land, M. F. (2005). Visual memory for objects in natural scenes: from fixations to object files. *The Quarterly journal of experimental psychology. A, Human experimental psychology*, 58(5), 931–960.

Teh, Y. G., Ping Pang, N. T., Shoesmith, W. D., Loo, J. L., & Teh, S. J. (2021). Team reflections: how co-created knowledge objects enhance knowledge integration. *Reflective Practice,* 22(4), 550-565.

Theeuwes, J. (2013). Feature-based attention: it is all bottom-up priming. *Phil Trans R Soc B*, 368: 20130055.

Thompson, S., Foulsham, T., Leekam, S. & Jones, C. (2019). Attention to the face is characterised by a difficult to inhibit first fixation to the eyes. *Acta Psychologica*. 193. 229-238.

Treisman, A. (1993). The perception of features and objects. In eds. Baddley, A & Weiskrantz, L. (1993). *Attention: Selection, Awareness, and Control. A Tribute to Donald Broadbent*. Clarendon Press, Oxford.

Treisman, A. & Schmidt, H. (1982). Illusory conjunctions in the perception of objects, *Cognitive Psychology*, 14(1), 107-141.

Tusher, M. (2022). Processing Speed and Comprehensibility of Visualizations and Texts. *Proceedings of CESCG 2022: The 26th Central European Seminar on Computer Graphics (non-peer-reviewed).*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. In: *Advances in neural information processing systems*. Red Hook: Curran Associates, 6000–6010.

Vaughan Van Hecke, A., Mundy, P., Block, J., Delgado, C., Parlade, M., Polmares, Y., and Hobson, J. (2012). Infant responding to joint attention, executive processes, and selfregulation in preschool children. *Infant Behav. Dev*. 35, 303–311.

Vecera, S. P., & Rizzo, M. (2006). Eye gaze does not produce reflexive shifts of attention: evidence from frontal-lobe damage. *Neuropsychologia*, 44(1), 150–159.

Villamor, M.M. & Rodrigo, M (2019). Gaze collaboration patterns of successful and unsuccessful programming pairs using cross-recurrence quantification analysis. *RPTEL* 14, 25.

Weldon, M. S., & Bellinger, K. D. (1997). Collective memory: Collaborative and individual processes in remembering. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23, 1160–1175.

Wheeler, S. & Passmore, J. (2020). All to play for: LEGO® SERIOUS PLAY® and its impact on team cohesion, collaboration and psychological safety in organisational settings using a coaching approach. *Journal of Work-Applied Management,* 12(2), 141-157.

Wiese, E., Wykowska, A., Zwickel, J. & Müller, H. J. (2012). I See What You Mean: How Attentional Selection Is Shaped by Ascribing Intentions to Others. *PLoS ONE* 7(9): e45391.

Wiley, M. & Wiley, J. F. (2019). Advanced R Statistical Programming and Data Models: Analysis, Machine Learning, and Visualization. 1st ed. 2019. *Berkeley, CA: Apress, 2019. Web.*

Williams, J. H., Waiter, G. D., Perra, O., Perrett, D. I., & Whiten, A. (2005). An fMRI study of joint attention experience. *NeuroImage*, 25(1), 133–140.

Wu, L., Bischof, D.W., & Kingstone, A. (2013). Looking while eating: The importance of social context to social attention. *Sci Rep* 3, 2356.

Yin, X. (2022). Influences of eye gaze cues on memory and its mechanisms: The function and evolution of social attention. *Front. Psychol*. 13:1036530.

Zenk, L., Hynek, N., Krawinkler, S. A., Peschl, M. F., & Schreder, G. (2021). Supporting innovation processes using material artefacts: Comparing the use of LEGO bricks and moderation cards as boundary objects. *Creativity and Innovation Management*, 30(4), 845–859.

Zhang, G., & Barmaki, R. (2020). Deep neural networks for collaborative learning analytics: Evaluating team collaborations using student gaze point prediction. *Australasian Journal of Educational Technology,* 36(6), 53-71.

APPENDIX 1

Structured long-term memory interview (in English)

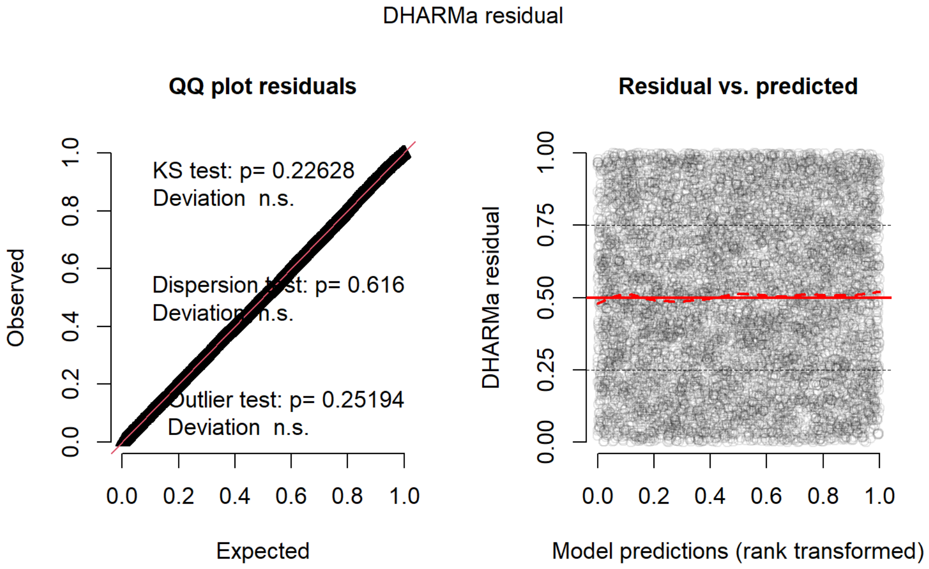
1. In the workshop, your task was to build a common understanding to the topic by utilizing objects. As a result of the workshop, a meaningful final build composed of objects was created.
2. What do you remember about your own contribution to the workshop's outcome (the final build)? We kindly ask you to share only your memories related to the final version of the build.
    a. Which objects in the final build were chosen by you?
    b. What meanings did you assign to your own objects?
    c. What meanings did you assign to the objects chosen by other participants?
3. What do you remember about the contributions of other participants to the workshop's outcome (the final build?) We kindly ask you to share only your memories related to the final version of the build.
    a. Which objects did other participants choose for the final build?
    b. What meanings did other participants assign to the objects in the final build?
4. Please describe the final build and its associated story as accurately as possible in your own words.

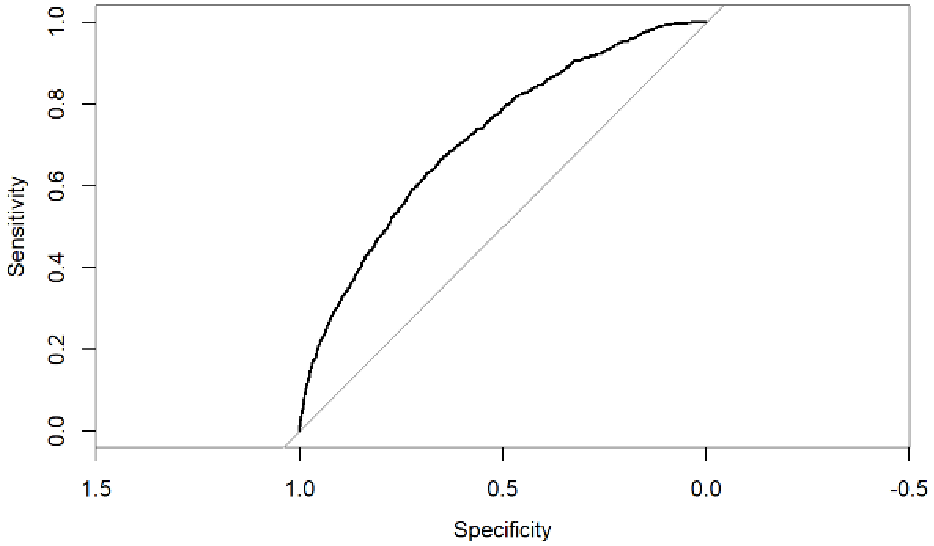
Strukturoitu jälkihaastattelu (Suomeksi)

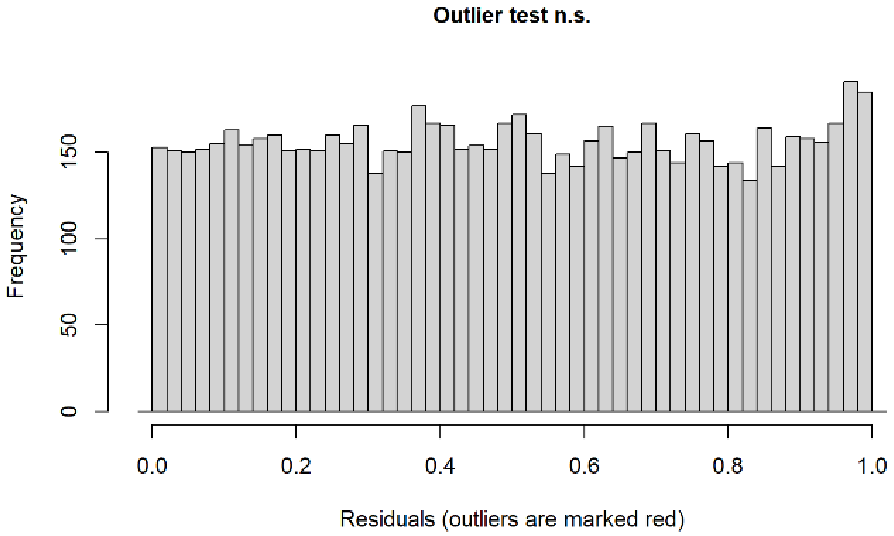
1. Työpajassa tehtävänne oli objekteja hyödyntämällä rakentaa yhteisymmärrys käsiteltävään aiheeseen. Työpajan lopputuloksena syntyi objekteista koottu merkityksellinen rakennelma.
2. Mitä muistat omasta osuudestasi rakennelman (objekteista tehty kokonaisuus) lopputulokseen? Pyydämme kertomaan vain rakennelman lopulliseen versioon liittyvät muistosi.
    a. Mitkä lopullisen rakennelman objektit olivat sinun valitsemia?
    b. Mitä merkityksiä annoit omille objekteille?
    c. Mitä merkityksiä annoit muiden osallistujien valitsemille objekteille?
3. Mitä muistat muiden osallistujien osuudesta rakennelman (objekteista tehty kokonaisuus) lopputulokseen? Pyydämme kertomaan vain rakennelman lopulliseen versioon liittyvät muistosi.
    a. Mitä objekteja muut osallistujat valitsivat lopputulokseen?
    b. Mitä merkityksiä muut osallistujat antoivat lopullisen rakennelman objekteille?
4. Kuvaile vielä lopullista rakennelmaa ja siihen liittyvää tarinaa mahdollisimman tarkasti omin sanoin.

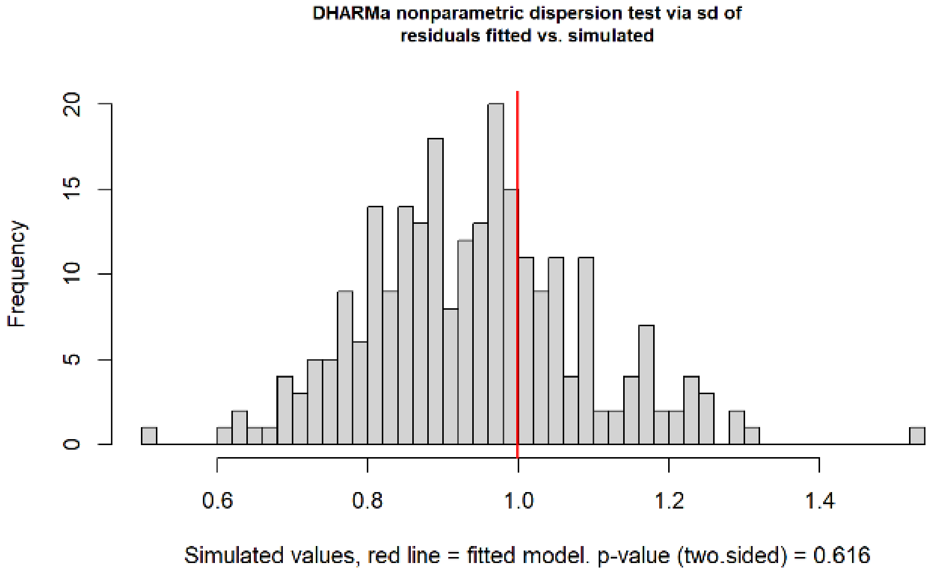
APPENDIX 2

Visual inspections of the model on initiating JVA and visual long-term memory.






Outlier test n.s.
Frequency
0
50
100
150
0.0
0.2
0.4
0.6
0.8
1.0
Residuals (outliers are marked red)


DHARMa nonparametric dispersion test via sd of residuals fitted vs. simulated
Frequency
0
5
10
15
20
0.6
0.8
1.0
1.2
1.4
Simulated values, red line = fitted model. p-value (two.sided) = 0.616