Does It MultiMatch? What Scanpath Comparison Tells us About Task Performance in Teams

Mohamad El Iskandarani, Jad A. Atweh, Shannon P. D. McGarry and Sara L. Riggs, University of Virginia, Charlottesville, VA, USA, **Nadine Marie Moacdieh**, Carleton University, Ottawa, Canada

Teamwork and collaboration form the cornerstones of organizational performance and success. It is important to understand how the attention allocation of team members is linked to performance. One approach to studying attention allocation in a team context is to compare the scanpath similarity of two people working in teams and to explore the link between scanpath similarity and team performance. In this study, participants were recruited to work in pairs on an unmanned aerial vehicle (UAV) task that included low and high workload conditions. An eye tracker was used to collect the eye movements of both participants in each team. The scanpaths of two teammates were compared in low and high workload conditions using MultiMatch, an established scanpath comparison algorithm. The obtained scanpath similarity values were correlated with performance measures of response time and accuracy. Several MultiMatch measures showed significant strong correlations across multiple dimensions, providing insight into team behavior and attention allocation. The results suggested that the more similar each team member's scanpath is, the better their performance. Additional research and consideration of experimental variables will be necessary to further understand how best to use MultiMatch for scanpath similarity assessment in complex domains.

Keywords: scanpath similarity, workload, team performance, eye tracking

Address correspondence to Mohamad El Iskandarani, System and Information Engineering Department, University of Virginia, 1827 University Avenue, Charlottesville 22903-1738, VA, USA.
Email: anz8av@virginia.edu
Journal of Cognitive Engineering and Decision Making
Vol. 0, No. 0, ■■ ■, pp. 1-16
DOI:10.1177/15553434231171484

Introduction

Teams are the foundation of many organizations and corporations, where a team is formally defined as two or more people who have precise roles and rely on one another to accomplish a common objective (Salas et al., 1992). From the ancient Greeks' army formations (Goldsworthy, 1997) to flight crews coordinating a 9,537-mile flight for nearly 19 hours (Pallini, 2020), teamwork and collaboration are at the core of various work environments. Researchers have studied team performance in different setups, including aviation (McNeese et al., 2018), military operations (Gorman et al., 2020; Meslec et al., 2020), and healthcare (Gorman et al., 2020).

A properly trained team can often achieve better results than one person alone and lead to a safer and more efficient system (Salas et al., 2008). However, working in teams in data-rich domains can also amplify the already-present complexity of operations, especially where human-machine interaction is involved (McNeese et al., 2018). In one instance, an unmanned aerial vehicle (UAV) crashed into the ground, with the accident later attributed to a lack of coordination between the operators handling the UAV (Williams, 2006).

Therefore, understanding what factors affect teamwork, and how this can be supported through display design, is an important human factors topic. This is especially important in complex, data-rich, and data-driven domains, where high mental workload can degrade team performance (Funke et al., 2012; Urban et al., 1995). Cognitive workload is defined as the gap between one's attentional resources and the cognitive demands placed on users (Wickens, 1992). Cognitive demands are typically varied by manipulating the user's task load

(Hancock et al., 1995) that is defined as the number of items that one has to attend to in order to successfully complete a task (Veltman & Gaillard, 1996).

However, it is still not clear how best to analyze the attention allocation of people working in teams (Atweh et al., 2022). There is a need for quantitative measures that can be used in real time and at a fine-grained level of analysis. This would allow for a better understanding of how people collaborate, which would, in turn, lead to better design principles for collaborative interfaces. Eye tracking is one approach that can be used in this regard, given that it can provide a trace of where a person is looking—the person's *scanpath*—which, in turn, can help shed further light on team performance (Devlin et al., 2020; Faulhaber & Friedrich, 2019). Scanpath analysis has been used in the past to assess mental workload (Maggi et al., 2019), cognitive capacity (Hayes & Henderson, 2017), task demand (Boot et al., 2009), and breast screening reading strategies (Chen et al., 2018), but these methods have not been implemented to assess team performance.

Thus, the overall goal of this study is to explore whether and to what extent the scanpath similarity of two people working together on a task is linked to their performance in a complex, multitasking, environment across different workloads. Scanpath similarity was assessed using the well-known scanpath comparison algorithm, MultiMatch (Dewhurst et al., 2012). We hypothesized that (1) pairs with more similar scanpaths would also have better performance (e.g., Maurer et al., 2018; Siirtola et al., 2019) and (2) the aforementioned relation would be accentuated during high workload periods. Understanding the link between scanpath similarity and performance can inform display design solutions and training instructions that ensure that teammates are effectively directing their attention as a function of workload. To this end, a simulator study in the context of UAV operations was conducted in which participants collaborated, working in pairs to complete multiple tasks that are akin to multi-UAV operations.

Background

Team Performance and Attention Allocation

Given the importance of good teamwork, several types of measures have been used to analyze the performance, perceived workload, and/or awareness of team members. Subjective questionnaires are one popular approach to gain insight on how operators thought they performed. One example of well-known questionnaires that have been used is the NASA Task Load Index (NASA-TLX; Hart & Staveland, 1988) that calculates workload scores based on several scales, such as mental demand, physical demand, temporal demand, performance, effort, and frustration. Another commonly used questionnaire is the Distributed Assessment of Team Mutual Awareness (DATMA), which measures mental workload and teamwork awareness (Berggren et al., 2011). Berggren et al. (2011) used both NASA-TLX and DATMA in their study to compare individual and team workload measures. Different types of performance measures have also been used for assessment, namely the response time to complete a task and the accuracy of the task performance. For instance, Jobidon et al. (2006) used the mean response time to detect a fire as a measure of team performance. Villamor and Rodrigo (2018), on the other hand, calculated a debugging score to assess teamwork among programmers. However, none of these measures provide insight into the attention allocation strategies of each team member, which can play an important role in understanding overall team performance as visual attention allocation is considered a multi-component cognitive resource that determines one's ability to focus and process information (Archibald et al., 2015).

One way to gain insight into attention allocation is by using eye tracking, an infrared-based technology that provides a trace of people's eye movements (Hess et al., 1998; Lin et al., 2004). Specifically, eye tracking provides output in terms of fixations and saccades. Fixations are spatially stable gaze points during which time visual processing takes place (Findlay, 2004). Saccades are the rapid eye movements in between fixations during which no visual

processing occurs (Yarbus, 1967). The sequence of fixations and saccades form the scanpath of each user and the areas that users look at on the screen are defined as areas of interest (AOIs). In recent years, eye tracking has been gaining interest as a means to assess and improve team performance (e.g., Daggett et al., 2017; D'Angelo & Begel, 2017; Devlin et al., 2020). Eye tracking provides a detailed, objective window into visual attention allocation and can also be used in real time (Lin et al., 2004). Understanding what teammates are looking at and when can provide new, previously unknown insight into their collective performance, as surveys, questionnaires, or debriefing strategies cannot precisely measure scanpath trends and their similarity.

Several eye tracking metrics have been used in the context of team performance. For example, pupillometry metrics have been applied to assess cognitive workload of people working in teams (e.g., Daggett et al., 2017). Other metrics that have been utilized include gaze overlap, which measures the times that several users are viewing the same area simultaneously (Pietinen et al., 2010). This was used by Devlin et al. (2020) to study the link between visual attention and pair performance during changes in workload. The phi coefficient (Φ; Bakeman & Gottman, 1997), which quantifies the lag between two time series, was also utilized by Devlin et al. (2020) as a measure of the coordination between scanpaths. In addition, cross recurrence (or gaze coupling or overlap) analysis has been used to measure how closely matched teammates' attention is, where cross recurrence occurs in general when two fixations from different people's scanpaths are within a certain radius of each other (Cherubini et al., 2010; Devlin et al., 2020; Jermann et al., 2010; Villamor & Rodrigo, 2018). Another notable approach is analyzing eye movement transitions using entropy-based statistical analysis (Krejtz et al., 2014), which can be applied to detect individual differences in eye movement transitions between AOI (Alemdag & Cagiltay, 2018).

The proposed theoretical underpinning for why shared cognition can improve performance is that it improves coordination and collaboration, which allows for more resources to be assigned to the task being performed (Langan-Fox et al., 2004). We thus centered our first hypothesis to reflect the notion that shared gaze will lead to better performance. For example, Brennan et al. (2008) found that team members whose gaze locations were very similar were twice as efficient in their searching tasks as solitary members. They were even more efficient than members who were able to talk together, a finding consistent with Neider et al.'s study (2010). D'Angelo and Begel (2017) developed a system where programmers were shown what the other was looking at while they worked, and they found providing this shared gaze information aids in coordination and effective communication. GazeTorch, a shared gaze interface developed by Akkil et al. (2016), was also found to make collaboration more effortless. Several other studies found that shared gaze improved performance and remote collaboration in teleconferencing (Gupta et al., 2016), video conferencing systems (Lee et al., 2017), problem solving (Schneider & Pea, 2013), collaborative visual search (Siirtola et al., 2019), and competitive and cooperative online gaming (Maurer et al., 2018).

On the other hand, Villamor and Rodrigo (2018) concluded that gaze recurrence alone was not a good predictor of pair success. For example, Müller et al. (2013) found that shared gaze in a puzzle solving task can induce uncertainty and delay. Another study by Zhang et al. (2017) concluded that shared gaze can potentially boost collaboration but can be impeded by factors such as trust and privacy. This uncertainty suggests that it is still not clear how shared attention is linked to task performance in teams. There is a need to explore other eye tracking metrics that may be able to better capture shared attention allocation.

Scanpath Comparison

One such measure could be scanpath comparison, which has not been explored to date in the context of teams. Such metrics may be able to provide additional insights regarding the link between the attention

allocation of teams and the performance of these teams. There are a number of algorithms that can provide a measure of scanpath similarity. ScanMatch is one notable example (Cristino et al., 2010). It has been used to compare the scanpaths of physics problem solvers (Madsen et al., 2012), discover the preferences of individuals with autism (Król & Król, 2020), and study complex visual search patterns (Frame et al., 2019). This method is based on the Needleman–Wunsch algorithm (Needleman & Wunsch, 1970) that was created to compare biological sequences. The ScanMatch algorithm includes two main steps: (1) creating sequences of letters that denote the sequence of AOIs fixated by the user and (2) calculating similarity scores between these sequences. The similarity score is a value ranging between 0 and 1. The higher the score, the more similar the scanpaths are; in other words, two identical sequences of AOIs would then result in a ScanMatch score of 1. ScanMatch thus provides a single quantitative measure of the similarity of two scanpaths. However, ScanMatch's dependence on AOIs means that AOI's size and order can greatly affect the output (Anderson et al., 2015). In addition, condensing scanpath similarity to just one measure does not paint a complete or detailed picture of what is going on; it neither provides insight into the duration of team member's scanpaths and how they are related nor in what aspects the scanpaths are similar. This illustrates the need for a multidimensional measure.

MultiMatch is one such scanpath comparison method that attempts to address some of the limitations of ScanMatch. It has been used in the literature in experiments to test memory performance (Foulsham et al., 2012), assess student cognitive processes (Stranc & Muldner, 2020), and study weather forecasters' decision-making processes (Wilson et al., 2018). MultiMatch is also notable for its robustness, as it manages spatial noise and perturbed scanpaths well (Dewhurst et al., 2012). In addition, the code for calculating MultiMatch is freely available online (Dewhurst et al., 2012). The MultiMatch algorithm requires a number of steps. First, the scanpath is converted into a series of vectors, each one representing a saccade (Dewhurst et al., 2012). The scanpath undergoes several simplifications (Figure 1). The first simplification consists of combining vectors of similar directions into one. Another simplification is amplitude based, in which consecutive saccades that have amplitudes less than a preset threshold are clustered into a single vector. Next, the scanpaths are temporally aligned (Dewhurst et al., 2012). Next, the corresponding vectors are compared (Foulsham et al., 2012). Five separate comparisons are performed, resulting in five measures (shape, length, direction, position, and duration). The results are averaged across the number of vectors and normalized to yield a value ranging between 0 and 1, where 1 represents perfect similarity. Each measure has its own significance and represents a certain spatial or temporal aspect of similarity as seen in Table 1 (Anderson et al., 2015). MultiMatch's different measures allow for assessing scanpath similarity at a more fine-grained level than ScanMatch, and it also allows for the comparison of scanpaths that have different lengths (Dewhurst et al., 2012). It is important to note that absolute scores of each MultiMatch measure cannot be compared against each other as each measure is calculated and normalized differently (Dewhurst et al., 2012; Wilson et al., 2018). Even though one downside of MultiMatch is that the threshold needs to be carefully selected, its present advantages were the reason it was selected for this study.

Method

Participants

Ten pairs of undergraduate students at the University of Virginia (20 students total) were recruited for the study (M = 21.3 years, SE = 0.24 years). Each pair consisted of one male and one female who did not previously know each other. The experiment lasted between 75 and 90 minutes and took place in a single session. Participants were compensated $10/hour for their time. This study was approved by the University of Virginia Institutional Review Board (protocol number 3480).

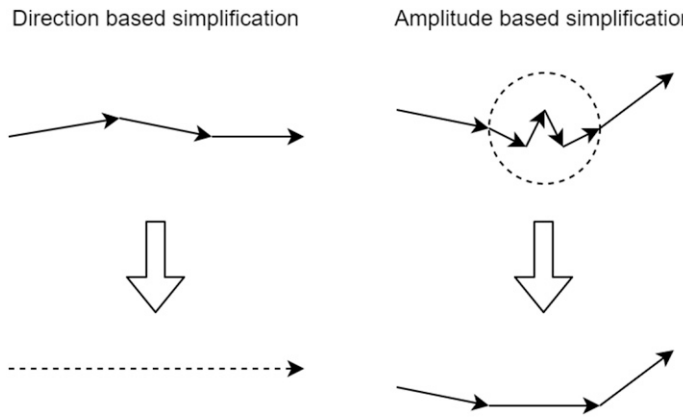
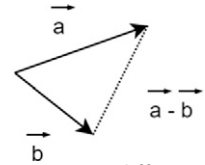
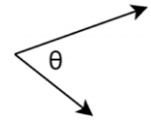
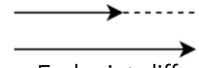
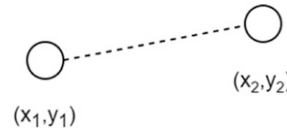
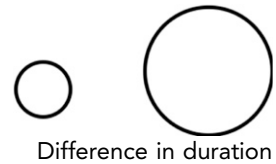


Figure 1. Illustration of the simplification steps.

Table 1. Summary of MultiMatch Measures and Their Indications (Anderson et al., 2015; Dewhurst et al., 2012; Wilson et al., 2018).

Measure	Definition
Shape similarity	 Vector difference between aligned saccade pairs
Direction similarity	 Angular difference between aligned saccades
Length similarity	 Endpoint difference in length of the pair
Position similarity	 Euclidean distance between aligned fixations
Duration similarity	 Difference in duration between aligned fixations

In all cases, a measure of 1 would indicate perfect similarity between the two scanpaths being compared.

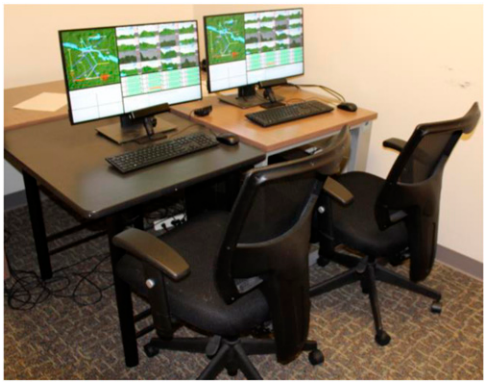


Figure 2. Experimental setup with the simulation shown on two networked computers.

Experimental Design

There were two workload conditions, low and high, that were manipulated by varying the number of active UAVs for the primary (target detection) task. For the low workload condition, 3–5 UAVs were active at all times, while 13–16 UAVs were active at all times for the high workload condition. These numbers were validated using NASA-TLX and performance measurements (see Devlin et al. (2020) for the full details). In each experimental condition, pairs completed two 15-minute trials, one with each of the two workload conditions. Pairs always completed the low workload condition before the high workload condition. The design of the simulation was based on the 'Vigilant Spirit Control Station' the Air Force uses to develop interfaces to control multiple UAVs (Feitshans et al., 2008). The simulation was developed using the Unity gaming engine and ran on a desktop computer (28" monitor, 2560 × 1440 screen resolution; Figure 2). Participants sat 26–28 inches from the monitor and used a standard mouse to input responses. Pairs were collocated, but each participant viewed separate monitors and used separate mice to input responses. The simulation was networked so participants could see inputs from their partner in real time (e.g., when Participant 1 responded to a chat message, Participant 2 could see his/her response in real time).

Two desktop-mounted FOVIO eye trackers with a sampling rate of 60 Hz were used to collect point of gaze data. The average degree of error for this eye tracker is 0.78° (SD = 0.59°). An external microphone was also used to record all verbal communication.

UAV Tasks and Point Values

Each pair was responsible for completing a primary task and three secondary tasks—that is, four tasks total—for up to 16 UAVs. Although all tasks were the pair's responsibility, only one participant from each pair had to complete each task. The four tasks were as follows:

1. *Target detection task (primary task).* Pairs monitored each UAV's video feed and indicated whether a target—a semi-transparent cube—was present. When a UAV was approaching a waypoint (predetermined area of interest denoted on the Map panel), its video feed could become "active" (i.e., video feed became highlighted; Figure 3). If a semi-transparent cube appeared while the video feed was active, the pair was instructed to press the target button to indicate a target was present; if no target was present, then no response was necessary. UAV video feeds were active for 10 seconds and a target could appear with 4–7 seconds left in this time interval. Pairs were instructed that the target detection task had the highest priority among the four tasks. In the low workload condition, one target appeared on one of the active UAV video feeds every 10 seconds. For the high workload condition, three targets appeared on three different active UAV video feeds every 10 seconds.
2. *Reroute task (secondary task).* Pairs were tasked to reroute a UAV when it was projected to enter a *no-fly-zone*, denoted by a red square on the Map panel (Figure 3). To reroute a UAV, a participant clicked on a respective UAV's numbered square in the Reroute Menu panel to activate the reroute menu that listed three alternative route options. Participants could click 'Preview' to see a specific alternative's suggested route. When the UAV was not rerouted in time (i.e., entered a *no-fly-zone*), it would no longer be able to complete the remainder of the mission. The rerouting task occurred 17 times in each condition.

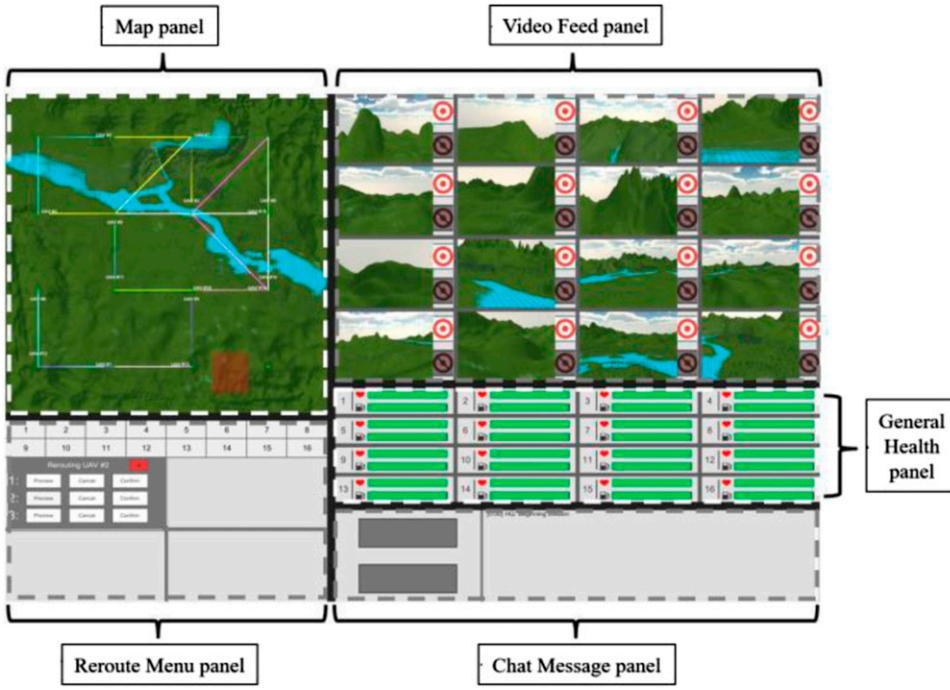


Figure 3. A Screenshot of the UAV simulation with panels labeled.

3. *Fuel leak task (secondary task).* Pairs were also tasked with monitoring and maintaining the overall health of each UAV. Participants used the General Health panel, that consisted of a *health status bar* and *fuel level bar* for each UAV (Figure 3). One instance where a UAV would need assistance is if it experienced a fuel leak, that consisted of the UAV's *fuel level bar* rapidly decreasing in fuel, the color of its *health status bar* changing from green to yellow, and the message "FIX LEAK" appearing in the *health status bar.* To stop a fuel leak, the participant clicked on the *health status bar.* This would change it back to green and stop the fuel from decreasing as rapidly. If the leak was not stopped in time, the UAV would reach the "FATAL FUEL LEAK" condition and the task could no longer be completed. A fuel leak occurred a total of 14 times for each condition.
4. *Chat message task (secondary task).* Pairs were tasked with responding to messages from headquarters by selecting one of two options on the left-hand side of the chat message panel (Figure 3). They were told to respond to as quickly and accurately as possible. There were 19 messages in each condition.

Table 2 shows the point value associated with each task (Devlin et al., 2019). Points were assigned to emphasize the priority of the primary task (i.e., target detection) as well as to convey the severity of incorrectly or not attending to a task (e.g., UAV flies through *no-fly-zone*). Also, we informed pairs that the highest scoring pair would earn an additional $10 to incentivize performance. Response times for each task for each pair were recorded as well.

Experiment Procedure

Participants read and signed the consent form and were then briefed about the experiment's goals and task expectations as a pair. Participants then independently completed a five-minute training phase. By the end of training, participants had to demonstrate they could achieve 70% accuracy for all tasks. We then informed the pairs about how the simulation was networked and provided them 3 minutes to introduce themselves to one another and discuss anything

Table 2. Point System for UAV Simulation.

Response to Task	Points Per Response
Correctly recognizing a target	+100
Correctly recognizing a non-target	+50
All secondary tasks (reroute, fuel leak, and chat message)	+30
Any incorrect or lack of response (false positive or negative to target detection task, UAV flies through *no-fly-zone*, or "FATAL FUEL LEAK" condition)	−100

they deemed necessary. There were no restrictions on how the participants could interact during these 3 minutes, that is, the experimenter gave no guidance on what should be discussed, so discussing team strategies before the experimental portion was completely participant-driven. Afterward, the audio recording started, and participants completed the low workload condition, were provided a short break, and then completed the high workload condition. Participants could communicate verbally with each other during the experiment. The same tasks appeared at both stations and the actions of each team member were reflected on both stations, but a participant could not see the cursor movements of their teammate. At the conclusion of the experiment, participants were compensated for their time.

Data Analysis

After we gathered the eye tracking data from the FOVIO eye tracker, we filtered the datasets and removed invalid entries. The data loss across all participants and trials was on average 11.9% (SD = 11.2%). We detected fixations and saccades using the code developed by Riggs Lab. This code is used to analyze eye tracking data collected from experimental studies with participants and it serves two main purposes: (1) filtering the eye tracking dataset and (2) detecting fixations and saccades based on Nyström and Holmqvist's (2011) velocity-based and data-driven adaptive algorithm. The code, implemented in Python, first takes the raw eye tracking files as input, and filters out empty or invalid recordings. Then, it passes the data through a Butterworth smoothing filter and calculates the angular velocities in preparation for the data-driven iterative algorithm that keeps iterating until the absolute difference between the newly calculated velocity threshold and the previous one converges to less than 1°.

We then used MATLAB to calculate the MultiMatch similarity scores for each pair of participants (one set of scores for low workload and another for high workload). The five measurements were extracted for each condition using the *doComparison* function, the main algorithm of MultiMatch (Dewhurst et al., 2012). The eye tracking data were then divided into one-minute segments and the algorithm was run for each segment in turn. Each pair thus had between 10 and 14 segments to run, and additional code was written to perform the *doComparison* function in batches. This process had to be done due to the large size of the data files that exceeded the RAM limit available. Note that, for our experiment, the *SimplifyExcel* function in the toolbox that pre-processes the eye tracking data was not used, as all the necessary pre-processing had been done beforehand by the event detection software.

Finally, we calculated the Pearson correlation coefficients between each MultiMatch measure (i.e., shape, length, position, direction, and duration) and each of the performance measures (points and response time). This was done for low workload and high workload separately, resulting in six Pearson correlation coefficients (and their associated p-value) per each low or high workload condition. The assumptions of normality (assessed using Shapiro-Wilks tests) were met for all variables, and homoscedasticity was checked using plots. In addition, Welch paired t-tests were used to compare the

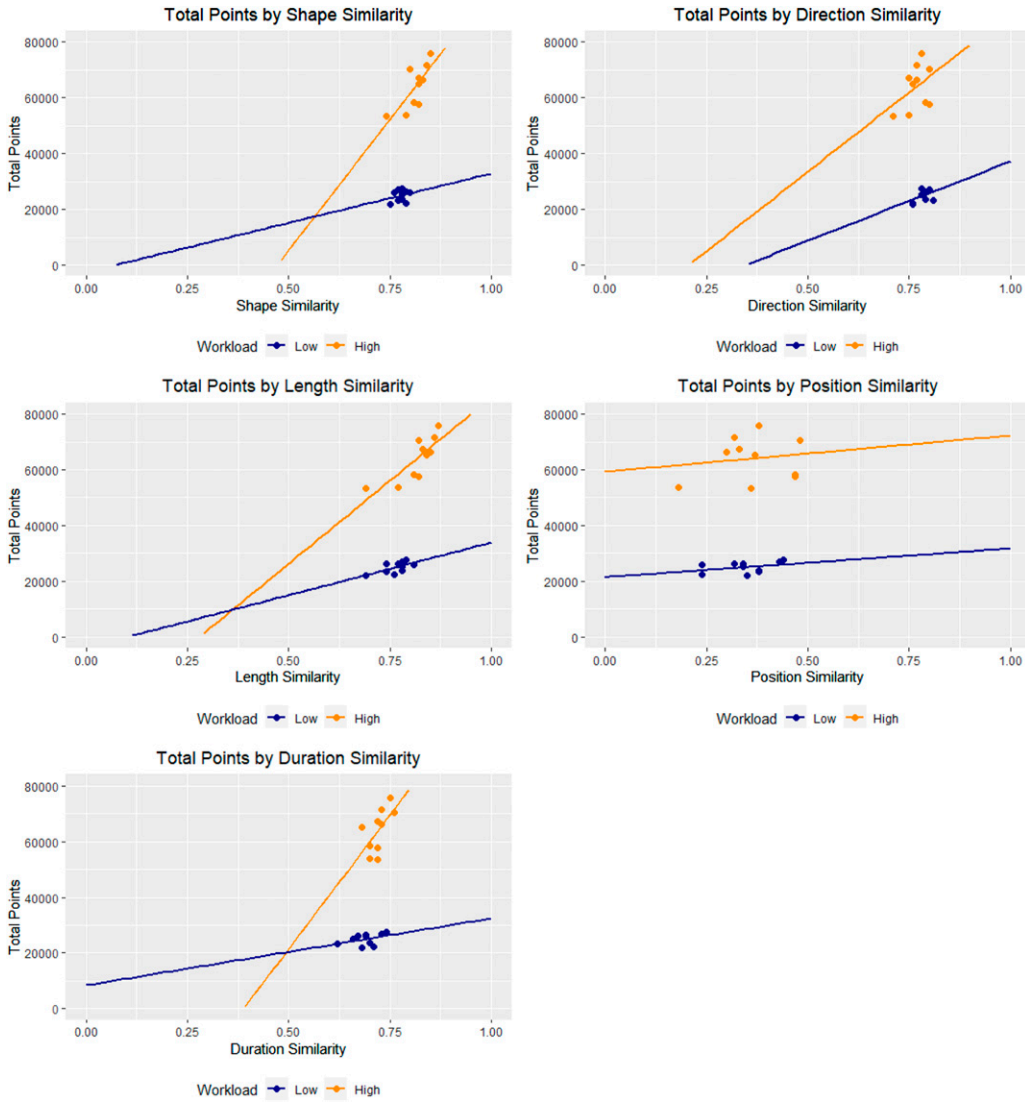


Figure 4. Scatter plots displaying total points by shape similarity, length similarity, direction similarity, position similarity, and duration similarity with best fitted lines at both low and high workloads.

performance results in low and high workload. These tests were used since the variances of the performance results at each workload condition were unequal. In all cases, significance was considered at $p < .05$.

Results

MultiMatch and Performance

Figure 4 shows the total points as a function of each of the MultiMatch measures (shape, direction, length, position, and duration) in both low and high workload conditions, where each point represents a team. Figure 5 shows the response time as a function of each of the MultiMatch measures for both workload conditions. Both Figures 4 and 5 contain best fitted lines for each workload condition. Table 3 shows the correlation values between MultiMatch values and the two measures of performance (points and response time).

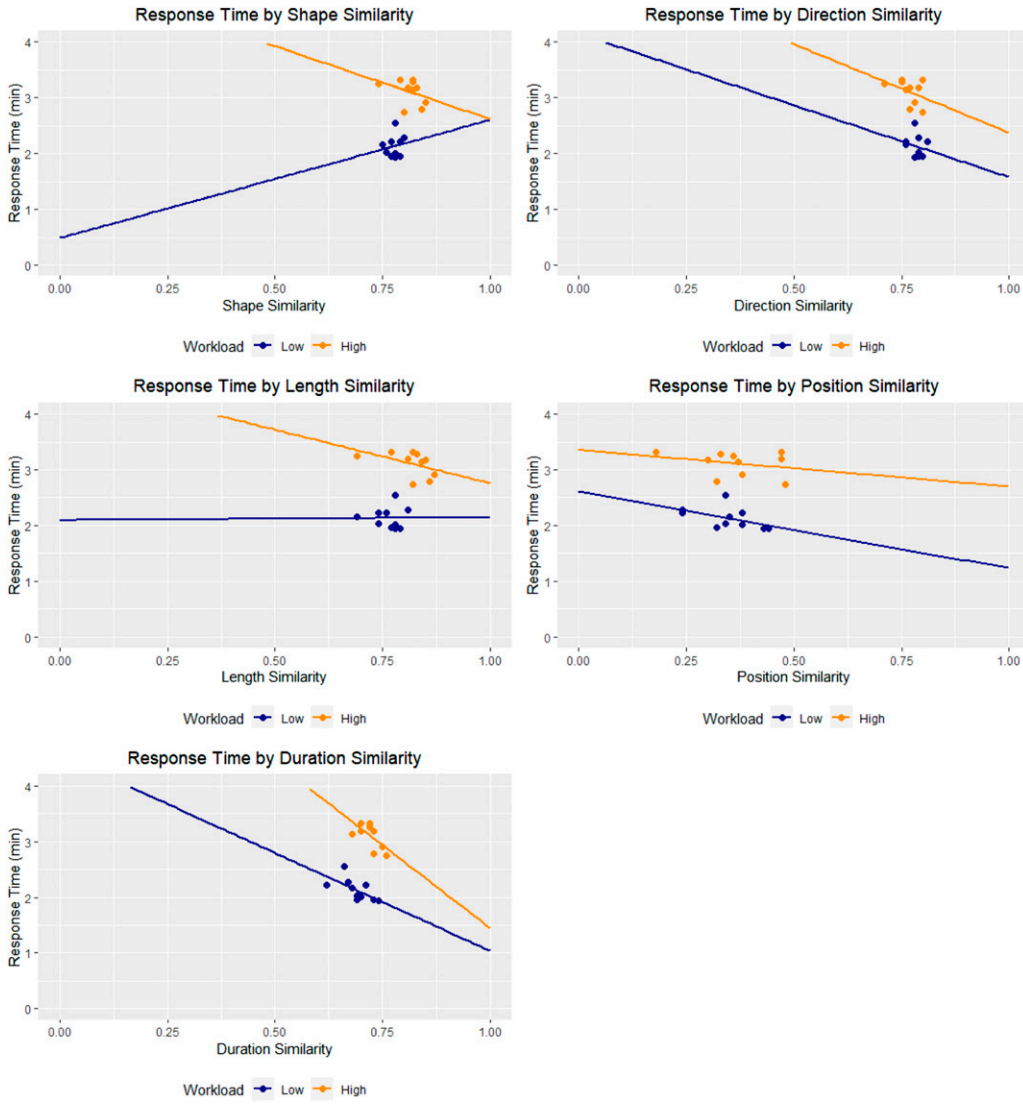


Figure 5. Scatter plots displaying response time (min) by shape similarity, length similarity, direction similarity, position similarity, and duration similarity with best fitted lines at both low and high workloads.

Workload Conditions

The mean of total points scored in the low workload condition was 24,942 (SD = 2,003) and for high workload it was 63,991 (SD = 7, 772). The mean response time in the low workload condition was 2.13 (SD = 0.194) and for high workload it was 3.11 (SD = 0.22). Welch paired t-tests revealed significant differences in total points ($t(9) = -19.24$, $p < .001$) and response time ($t(9) = -16.51$, $p < .001$) means between low and high workload. In addition, the correlation coefficients of Table 3 that showed significance (i.e., MultiMatch measures of shape, length, and duration correlated with each of points and response time) during low workload were Fisher Z transformed (Fisher, 1915) and z-tested against the high workload condition. In the case of points, there was a significant difference in the means of the two samples ($z = -2.26$, $p = .024$), while for response time the difference was not significant ($z = 1.2$, $p = 0.22$).

Table 3. Correlation Analysis Between the Five MultiMatch Measures (Shape, Direction, Length, Position, and Duration) and the Two Performance Measures (Total Points and Response Times) for Low and High Workload.

		Low Workload		High Workload	
MultiMatch Measure Similarity	Dependent Variable	*Correlation Coefficient*	*p-Value*	*Correlation Coefficient*	*p-Value*
Shape	Total points	0.26	.46	**0.75**	<.005
	Response time	0.16	0.651	−0.37	0.29
Direction	Total points	0.45	0.19	0.4	0.25
	Response time	−0.21	0.56	−0.39	0.25
Length	Total points	**0.64**	.048	**0.81**	<.005
	Response time	.008	0.98	−0.46	0.18
Position	Total points	0.36	0.30	0.16	0.67
	Response time	−0.48	0.16	−0.28	0.43
Duration	Total points	0.41	0.20	0.59	0.07
	Response time	**−0.63**	.005	**−0.65**	<.005

Values in bold represent strong significant correlations (i.e., an absolute correlation value above 0.6; Jurs et al., 1998).

Discussion

The overall goal of this experiment was to analyze whether and to what extent the scanpath similarity of two people working together on a complex task was indicative of team performance, and whether this differed during low and high workload. We had hypothesized that: (1) pairs with more similar scanpaths would have better performance and (2) these performance benefits will be accentuated during high workload.

Hypothesis 1: Pairs with more similar scanpaths have better performance

We had predicted that a higher similarity between participants' scanpaths would result in better performance. In other words, if the participants' attention allocation strategies were similar in terms of location, shape, sequence, etc., they would be assumed to be more synchronized and more aware of each other's actions. This would, in turn, enable better team performance. This would then translate to a positive correlation between scanpath similarity and total points, and a negative correlation between scanpath similarity and response time (faster response times meant better performance). This would be in line with previous work that showed pairs with similar attention allocation performed better as a team (e.g., Cherubini et al., 2010; D'Angelo & Begel, 2017), albeit without the level of detail provided by MultiMatch. The findings here could also extend to build on the literature on gaze sharing, that is, allowing teams to view each other's gaze points on their respective displays while simultaneously completing their tasks, which has been shown to improve performance (Lee et al., 2017).

Our hypothesis held true for two dimensions of the MultiMatch algorithm: length similarity and duration similarity. For length similarity, there was a strong (>0.6) and significant positive correlation with total points both in low and high workload, while for duration similarity there was a strong negative correlation with response time in both low and high workload. This suggests that similarities in teammates' saccade lengths and fixation durations matter more than similarities in their fixation positions. It appears that *how* teammates scan makes more of an impact than *where* exactly the pair was looking, as evidenced by the low and non-significant correlation for position similarity. It is important to emphasize that what matters here is not necessarily the saccade length or fixation duration of each team member per se, but rather that these are similar for both teammates. Similarly, the high correlation coefficients for

length similarity and total points suggest that similarity in saccade length indicates better team performance as well.

Hypothesis 2: Performance benefits are accentuated during high workload

The significant difference in performance measures between the low and high workload conditions confirms that performance decrements did occur due to the workload manipulation. We expected there would be a stronger link between scanpath similarity and performance during high workload, that is, more positive correlation coefficients with total points scored and more negative ones with response time. This was observed for three of the five measures: shape, length, and duration. These measures showed higher correlation coefficients (in absolute value) than their low workload counterpart. The effect of workload was also evident in the significant Fisher z-test results and the slopes of the best fitted lines, where the high workload slopes were greater than their low workload counterparts for all five MultiMatch measures (shape, direction, position, length, and duration similarity). We posit several explanations for these results. First, in the more challenging high workload condition, there is a stronger correlation between scanpath similarity and total points scored. This may be due to the teammates becoming more focused on the task that modulated workload and resulted in the team narrowing their attention allocation to the respective AOI (as evident in Devlin et al., 2019). This could explain why the teams had more similar scanpaths (Wickens & Alexander, 2009). This was true of the best performing pairs as they had a change in their attention allocation strategy, that is, having a more focused strategy that resulted in more similar scanpaths during high workload compared to a more open-ended/free-gaze strategy during low workload.

Second, a notably high correlation coefficient was between shape similarity and total points scored in the high workload condition that indicates that team members who had more similar scanpath shapes performed better. Dewhurst et al. (2012) noted that shape similarity has been found to be important in fields such as visual imagery research, where fixation order and position are not as crucial as in interfaces that have very clear-cut and well-structured AOIs, such as a website. For example, Gbadamosi and Zangemeister (2001) used scanpath shape to compare scanpaths when participants were viewing an image. Given the present testbed consisted of a complex interface with a lot of imagery (e.g., the video feeds), this may be the reason for the observed relation. It seems that shape similarity is capturing a unique and specific aspect of teammates' scanpaths and therefore it may be a valid indicator of team performance in a visually data-rich environment.

Thus, it appears that shape, length, and duration similarity are the aspects of MultiMatch that are best suited to assess the performance of teams experiencing high workload in complex domains, much like this experiment's simulation. It could be that position and direction similarity will be more strongly linked to performance in the context of a simpler/more directed task with fewer areas and targets that can be carefully defined using AOIs.

Overall, MultiMatch appears to be a useful and very promising tool for assessing team attention allocation strategies and how they related to performance especially during high workload periods. The strong correlation of performance with the three MultiMatch measures (shape, length, and duration) can help provide suggestions for interface design and teamwork strategies in complex, multitasking domains. The findings provide support for developing training programs that teach teammates how to coordinate their scanpaths as a means to optimize team performance. This could be done by showing novice teams the scanning approach of expert teams. For example, novices who were trained to mimic expert's visual patterns while reading medical images of lungs (Dempere-Marco et al., 2002) or chest X-rays (Litchfield et al., 2010) showed improved performance. The findings also provide support for design solutions that encourage teammates to scan a display in a similar fashion. For example, the system could highlight what a team member is looking

at/scanning (e.g., a box changes color to highlight the shared area if both users are looking at the same lines of code; D'Angelo & Begel, 2017). These developments would be especially beneficial in high workload and data-rich settings, such as emergency dispatching or process control.

Limitations and Future Work

Overall, our MultiMatch values were similar to those of Foulsham et al. (2012), with the exception of length similarity, where our values were generally lower. By definition, length similarity is the absolute amplitude difference of aligned saccades, so the nature of the task and the display layout may impact this measure. For example, a task that involves navigating rapidly between different sections of the screen like in our experiment might yield different length similarity values than a task of focusing on a static image or object. It is thus important not to use just one measure of MultiMatch when assessing team performance and to always consider a team's context when generalizing results. Additional studies in different contexts are needed to improve the external validity of the experiment. Future studies could also control and/or analyze other aspects of team collaboration, such as the communication between team members. Another limitation of our experiment was the small sample size of 10 teams, whereas a larger sample size may have yielded more significant correlations (like the marginal correlation of duration with total points at high workload that has a value of 0.59 and a p-value of 0.07).

Future work can further explore MultiMatch as a scanpath comparison tool by implementing it across different types of domains and tasks, for example, the pair programming collaboration setup in Villamor and Rodrigo (2018). It would be interesting to see how many of the same conclusions hold true for different types of tasks, contexts, and performance measures. For example, system failures could be integrated to investigate how teams adapt to unexpected events and tasks. Also, the results of the current experiment can be used to inform the design of human-robot/artificial intelligence teams; for instance, it would be interesting to investigate the effect of variables like agent autonomy and team composition (O'Neill et al., 2022) on team performance and if the effect can be captured using scanpath similarity measures. Furthermore, an interesting future research direction would be to study the effect of pre-experiment communication on the scanpath similarity and performance of teams. In other words, if the teammates agree to a certain strategy, such as attending to mutually exclusive tasks on separate parts of the screen, it would be interesting to see if that would lead to better performance over time. There is also merit in analyzing how the team communication impacts scanpath similarity and team performance and whether the trends evolve over time. It would also be worth studying the role different personalities (e.g., De Raad, 2000) have in the currently observed relationship between scanpath similarity and task performance. Also, we could explore the effect the point system had on current results, motivating participants, and informing strategy. Conversely, we could explore whether removing any motivating factor would reduce the competitive edge and lead to "social loafing," that is the decrease in efforts exerted by the individual when working in a group setting (Liden et al., 2004). If MultiMatch metrics could capture the latter, this would be very informative and impactful in complex, dynamic domains.

Acknowledgments

This study was supported in part by the National Science Foundation (NSF grant: #2008680; Program Manager: Dr. Dan Cosley). The authors would also like to thank Aakash Bhagat for the development of the simulator used in this study.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ORCID iDs

Mohamad El Iskandarani https://orcid.org/0000-0002-2779-474X

Jad A. Atweh https://orcid.org/0000-0001-6657-0792
Shannon P. D. McGarry https://orcid.org/0000-0003-2652-012X
Sara L. Riggs https://orcid.org/0000-0002-0112-9469
Nadine Marie Moacdieh https://orcid.org/0000-0002-7677-7946

References

Akkil, D., James, J. M., Isokoski, P., & Kangas, J. (2016). GazeTorch: Enabling gaze awareness in collaborative physical tasks, Proceedings of the 2016 CHI conference extended Abstracts on human factors in computing systems (pp. 1151–1158). Association for Computing Machinery. https://doi.org/10.1145/2851581.2892459

Alemdag, E., & Cagiltay, K. (2018). A systematic review of eye tracking research on multimedia learning. *Computers & Education*, *125*, 413–428. https://doi.org/10.1016/j.compedu.2018.06.023

Anderson, N. C., Anderson, F., Kingstone, A., & Bischof, W. F. (2015). A comparison of scanpath comparison methods. *Behavior Research Methods*, *47*(4), 1377–1392. https://doi.org/10.3758/s13428-014-0550-3

Archibald, L. M. D., Levee, T., & Olino, T. (2015). Attention allocation: Relationships to general working memory or specific language processing. *Journal of Experimental Child Psychology*, *139*, 83–98. https://doi.org/10.1016/j.jecp.2015.06.002

Atweh, J. A., Marie Moacdieh, N., & Riggs, S. L. (2022). Identifying individual-, team-, and organizational-level factors that affect team performance in complex domains based on recent literature. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 66, No. 1, pp. 1795–1799). Sage Publications. https://doi.org/10.1177/1071181322661213

Bakeman, R., & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511527685

Berggren, P., Prytz, E., Johansson, B., & Nahlinder, S. (2011). The relationship between workload, teamwork, situation awareness, and performance in teams: a microworld study. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *55*(1), 851–855. https://doi.org/10.1177/1071181311551177

Boot, W. R., Becic, E., & Kramer, A. F. (2009). Stable individual differences in search strategy?: the effect of task demands and motivational factors on scanning strategy in visual search. *Journal of Vision*, *9*(3), 7.1–716. https://doi.org/10.1167/9.3.7

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. https://doi.org/10.1016/j.cognition.2007.05.012

Chen, Y., Gale, A., Dong, L., Tang, Q., & Bernardi, D. (2018). Analysis of visual search behaviour from experienced radiologists interpreting digital breast tomosynthesis (dbt) images: a pilot study. In R. M. Nishikawa & F. W. Samuelson (eds), *Medical imaging 2018: Image perception, observer performance, and technology assessment*. SPIE. https://doi.org/10.1117/12.2293615

Cherubini, M., Nüssli, M.-A., & Dillenbourg, P. (2010). This is it ! : Indicating and looking in collaborative work at distance. *Journal of Eye Movement Research*, *3*(5). https://doi.org/10.16910/jemr.3.5.3

Cristino, F., Mathôt, S., Theeuwes, J., & Gilchrist, I. D. (2010). ScanMatch: A novel method for comparing fixation sequences. *Behavior Research Methods*, *42*(3), 692–700. https://doi.org/10.3758/BRM.42.3.692

Daggett, M., O'Brien, K., Hurley, M., & Hannon, D. (2017). Predicting team performance through human behavioral sensing and quantitative workflow instrumentation. In I. L. Nunes (Ed.), *Advances in human factors and system interactions* (497, pp. 245–258). Springer International Publishing. https://doi.org/10.1007/978-3-319-41956-5_22

D'Angelo, S., & Begel, A. (2017). Improving communication between pair programmers using shared gaze awareness, Proceedings of the 2017 CHI conference on human factors in computing systems (pp. 6245–6290). https://doi.org/10.1145/3025453.3025573

Dempere-Marco, L., Hu, X.-P., MacDonald, S. L. S., Ellis, S. M., Hansell, D. M., & Yang, G. Z. (2002). The use of visual search for knowledge gathering in image decision support. *IEEE Transactions on Medical Imaging*, *21*(7), 741–754. https://doi.org/10.1109/TMI.2002.801153

De Raad, B. (2000). *The big five personality factors: The psycholexical approach to personality* (p. 128). Hogrefe & Huber Publishers.

Devlin, S. P., Flynn, J. R., & Riggs, S. L. (2019). Examining the visual attention of pairs of operators during a low to high workload change. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *63*(1), 201–205. https://doi.org/10.1177/1071181319631168

Devlin, S. P., Moacdieh, N. M., Wickens, C. D., & Riggs, S. L. (2020). Transitions between low and high levels of mental workload can improve multitasking performance. *IISE Transactions on Occupational Ergonomics and Human Factors*, *8*(2), 72–87. https://doi.org/10.1080/24725838.2020.1770898

Dewhurst, R., Nyström, M., Jarodzka, H., Foulsham, T., Johansson, R., & Holmqvist, K. (2012). It depends on how you look at it: scanpath comparison in multiple dimensions with multimatch, a vector-based approach. *Behavior Research Methods*, *44*(4), 1079–1100. https://doi.org/10.3758/s13428-012-0212-2

Faulhaber, A. K., & Friedrich, M. (2019). *Eye-tracking metrics as an indicator of workload in commercial single-pilot operations* (pp. 213–225).

Feitshans, G., Rowe, A., Davis, J., Holland, M., & Berger, L. (2008). Vigilant spirit control station (vscs): the face of counter aiaa guidance, navigation and control conference and exhibit. *AIAA guidance, navigation and control conference and exhibit*. https://doi.org/10.2514/6.2008-6309

Findlay, J. M. (2004). Eye scanning and visual search. In J. M. Henderson & F. Ferreira (Eds.), *The Interface of language, vision, and action: Eye Movements and the visual world* (pp. 134–159).

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, *10*(4), 507–521. https://doi.org/10.2307/2331838

Foulsham, T., Dewhurst, R., Nyström, M., Jarodzka, H., Johansson, R., Underwood, G., & Holmqvist, K. (2012). Comparing scanpaths during scene encoding and recognition: a multi-dimensional approach. *Journal of Eye Movement Research*, *5*(4). https://doi.org/10.16910/jemr.5.4.3

Frame, M. E., Warren, R., & Maresca, A. M. (2019). Scanpath comparisons for complex visual search in a naturalistic environment. *Behavior Research Methods*, *51*(3), 1454–1470. https://doi.org/10.3758/s13428-018-1154-0

Funke, G. J., Knott, B. A., Salas, E., Pavlas, D., & Strang, A. J. (2012). Conceptualization and measurement of team workload: a critical need. *Human Factors*, *54*(1), 36–51. https://doi.org/10.1177/0018720811427901

Gbadamosi, J., & Zangemeister, W. H. (2001). Visual imagery in hemianopic patients. *Journal of Cognitive Neuroscience*, *13*(7), 855–866. https://doi.org/10.1162/089892901753165782

Goldsworthy, A. K. (1997). The othismos, myths and heresies: the nature of hoplite battle. *War in History*, *4*(1), 1–26. https://doi.org/10.1191/096834497667100325

Gorman, J. C., Grimm, D. A., Stevens, R. H., Galloway, T., Willemsen-Dunlap, A. M., & Halpin, D. J. (2020). Measuring real-time team cognition during team training. *Human Factors*, *62*(5), 825–860. https://doi.org/10.1177/0018720819852791

Gupta, K., Lee, G. A., & Billinghurst, M. (2016). Do you see what i see? the effect of gaze tracking on task space remote collaboration. *IEEE Transactions on Visualization and Computer Graphics*, *22*(11), 2413–2422. https://doi.org/10.1109/TVCG.2016.2593778

Hancock, P., Williams, G., Manning, C., & Miyake, S. (1995). Influence of task demand characteristics on workload and performance. *The International Journal of Aviation Psychology*, *5*(1), 63–86. https://doi.org/10.1207/s15327108ijap0501_5

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology*, *52*, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

Hayes, T. R., & Henderson, J. M. (2017). Scan patterns during real-world scene viewing predict individual differences in cognitive capacity. *Journal of Vision*, *17*(5), 23. https://doi.org/10.1167/17.5.23

Hess, S., Ray, W., Goldberg, J., Hettinger, L., & Hass, M. (1998). Driving system adaptation: disambiguating non-invasive physiological measures with cognitive task analysis. *Psychological issues in the design and use of adaptive, virtual interface* (pp. 339–450). Lawrence Erlbaum Publishers.

Jermann, P., Nüssli, M.-A., & Li, W. (2010). Using dual eye-tracking to unveil coordination and expertise in collaborative tetris. *Electronic Workshops in Computing*. https://doi.org/10.14236/ewic/HCI2010.7

Jobidon, M.-E., Breton, R., Rousseau, R., & Tremblay, S. (2006). Team response to workload transition: the role of team structure. https://doi.org/10.1177/154193120605001710

Jurs, H. W., Hinkle, D., & Wiersma, W. (1998). *Applied statistics for the behavioral sciences*.

Krejtz, K., Szmidt, T., Duchowski, A. T., & Krejtz, I. (2014). Entropy-based statistical analysis of eye movement transitions. *Proceedings of the Symposium on Eye Tracking Research and Applications*, 159–166. https://doi.org/10.1145/2578153.2578176

Król, M. E., & Król, M. (2020). Scanpath similarity measure reveals not only a decreased social preference, but also an increased nonsocial preference in individuals with autism. *Autism: The International Journal of Research and Practice*, *24*(2), 374–386. https://doi.org/10.1177/1362361319865809

Langan-Fox, J., Anglim, J., & Wilson, J. R. (2004). Mental models, team mental models, and performance: process, development, and future directions. *Human Factors and Ergonomics in Manufacturing*, *14*(4), 331–352. https://doi.org/10.1002/hfm.20004

Lee, G. A., Kim, S., Lee, Y., Dey, A., Piumsomboon, T., Norman, M., & Billinghurst, M. (2017). Improving Collaboration in Augmented Video Conference using Mutually Shared Gaze, ICAT-EGVE 2017 - International conference on artificial Reality and Telexistence and Eurographics Symposium on Virtual environments (*8*). https://doi.org/10.2312/EGVE.20171359

Liden, R. C., Wayne, S. J., Jaworski, R. A., & Bennett, N. (2004). Social loafing: a field investigation. *Journal of Management*, *30*(2), 285–304. https://doi.org/10.1016/j.jm.2003.02.002

Lin, Y., Zhang, W. J., & Koubek, R. J. (2004). Effective attention allocation behavior and its measurement: a preliminary study. *Interacting with Computers*, *16*(6), 1195–1210. https://doi.org/10.1016/j.intcom.2004.08.006

Litchfield, D., Ball, L. J., Donovan, T., Manning, D. J., & Crawford, T. (2010). Viewing another person's eye movements improves identification of pulmonary nodules in chest X-ray inspection. *Journal of Experimental Psychology. Applied*, *16*(3), 251–262. https://doi.org/10.1037/a0020082

Madsen, A., Larson, A., Loschky, L., & Rebello, N. S. (2012). Using ScanMatch scores to understand differences in eye movements between correct and incorrect solvers on physics problems. *Proceedings of the Symposium on Eye Tracking Research and Applications*, *'12*, 193. https://doi.org/10.1145/2168556.2168591

Maggi, P., Ricciardi, O., & Di Nocera, F. (2019). Ocular indicators of mental workload: a comparison of scanpath entropy and fixations clustering. In L. Longo & M. C. Leva (eds), *Human mental workload: Models and applications* (1107, pp. 205–212). Springer International Publishing. https://doi.org/10.1007/978-3-030-32423-0_13

Maurer, B., Lankes, M., & Tscheligi, M. (2018). Where the eyes meet: lessons learned from shared gaze-based interactions in cooperative and competitive online games. *Entertainment Computing*, *27*, 47–59. https://doi.org/10.1016/j.entcom.2018.02.009

McNeese, N. J., Demir, M., Cooke, N. J., & Myers, C. (2018). Teaming with a synthetic teammate: insights into human-autonomy teaming. *Human Factors*, *60*(2), 262–273. https://doi.org/10.1177/0018720817743223

Meslec, N., Duel, J., & Soeters, J. (2020). The role of teamwork on team performance in extreme military environments: an empirical study. *Team Performance Management: An International Journal*, *26*(5/6), 325–339. https://doi.org/10.1108/tpm-02-2020-0009

Müller, R., Helmert, J. R., Pannasch, S., & Velichkovsky, B. M. (2013). Gaze transfer in remote cooperation: is it always helpful to see what your partner is attending to? *Quarterly Journal of Experimental Psychology*, *66*(7), 1302–1316. https://doi.org/10.1080/17470218.2012.737813

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*(3), 443–453. https://doi.org/10.1016/0022-2836(70)90057-4

Neider, M. B., Chen, X., Dickinson, C. A., Brennan, S. E., & Zelinsky, G. J. (2010). Coordinating spatial referencing using shared gaze. *Psychonomic Bulletin & Review*, *17*(5), 718–724. https://doi.org/10.3758/PBR.17.5.718

Nyström, M., & Holmqvist, K. (2010). An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. *Behavior Research Methods*, *42*(1), 188–204. https://doi.org/10.3758/BRM.42.1.188

O'Neill, T., McNeese, N., Barron, A., & Schelble, B. (2022). Human–autonomy teaming: a review and analysis of the empirical literature. *Human Factors*, *64*(5), 904–938. https://doi.org/10.1177/0018720820960865

Pallini, T. (2020). *Inside the new world's longest FLIGHT: What it's like to fly on Singapore airlines' new route between Singapore and New York*. https://www.businessinsider.com/inside-the-new-worlds-longest-flight-singapore-airlines-a350-tour-2020-11

Pietinen, S., Bednarik, R., & Tukiainen, M. (2010). Shared visual attention in collaborative programming: a descriptive analysis. *Proceedings of the 2010 ICSE Workshop on Cooperative and Human Aspects of Software Engineering*, 21–24. https://doi.org/10.1145/1833310.1833314

Salas, E., Cooke, N. J., & Rosen, M. A. (2008). On teams, teamwork, and team performance: discoveries and developments. *Human Factors*, *50*(3), 540–547. https://doi.org/10.1518/001872008X288457

Salas, E., Dickinson, T. L., Converse, S. A., & Tannenbaum, S. I. (1992). Toward an understanding of team performance and training. In R. W. Swezey & E. Salas (eds), *Teams: Their training and performance*. Ablex Publishing.

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. https://doi.org/10.1007/s11412-013-9181-4

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. https://doi.org/10.1080/10447318.2019.1565746

Stranc, S., & Muldner, K. (2020). Scanpath analysis of student attention during problem solving with worked examples. In I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (eds), *Artificial intelligence in education* (12164, pp. 306–311). Springer International Publishing. https://doi.org/10.1007/978-3-030-52240-7_56

Urban, J. M., Weaver, J. L., Bowers, C. A., & Rhodenizer, L. (1995). Effects of workload and structure on team processes and performance: implications for complex team decision making. *Human factors*, *11*.

Veltman, J., & Gaillard, A. (1996). Physiological indices of workload in a simulated flight task. *Biological Psychology*, *42*(3), 323–342. https://doi.org/10.1016/0301-0511(95)05165-1

Villamor, M., & Rodrigo, Ma. M. (2018). *Predicting successful collaboration in a pair programming eye tracking experiment* (pp. 263–268). Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization. https://doi.org/10.1145/3213586.3225234

Wickens, C. D. (1992). Workload and situation awareness: an analogy of history and implications. *Insight: The Visual Performance Technical Group Newsletter*, *14*(4), 1–3.

Wickens, C. D., & Alexander, A. L. (2009). Attentional tunneling and task management in synthetic vision displays. *The International Journal of Aviation Psychology*, *19*(2), 182–199. https://doi.org/10.1080/10508410902766549

Williams, K. W. (2006). 8. Human factors implications of unmanned aircraft accidents: flight-control problems. In *Human factors of remotely operated vehicles*. Emerald Group Publishing Limited.

Wilson, K. A., Heinselman, P. L., & Kang, Z. (2018). Comparing forecaster eye movements during the warning decision process. *Weather and Forecasting*, *33*(2), 501–521. https://doi.org/10.1175/WAF-D-17-0119.1

Yarbus, A. L. (1967). Eye movements during perception of complex objects. In A. L. Yarbus (Ed.), *Eye movements and vision* (pp. 171–211). Springer US. https://doi.org/10.1007/978-1-4899-5379-7_8

Zhang, Y., Pfeuffer, K., Chong, M. K., Alexander, J., Bulling, A., & Gellersen, H. (2017). Look together: using gaze for assisting co-located collaborative search. *Personal and Ubiquitous Computing*, *21*(1), 173–186. https://doi.org/10.1007/s00779-016-0969-x

Mohamad El Iskandarani is a PhD Candidate in Systems Engineering in the Department of Systems and Information Engineering at the University of Virginia. He received his M.E. in Systems Engineering from the University of Virginia in 2023 and his B.E. in Computer and Communications Engineering from the American University of Beirut in 2021.

Jad A. Atweh is a PhD Candidate in Systems Engineering in the Department of Systems and Information Engineering at the University of Virginia. He received his M.E. in Systems Engineering from the University of Virginia in 2023 and his B.E. in Industrial Engineering from the American University of Beirut in 2021.

Shannon P. D. McGarry is a General Engineer at Naval Research Laboratory. She earned a Ph.D. in Systems Engineering from the University of Virginia in 2021.

Sara L. Riggs is an Associate professor in Systems Engineering in the Department of Engineering Systems and Environment at the University of Virginia. She obtained her PhD in Industrial and Operations Engineering from the University of Michigan, Ann Arbor in 2014.

Nadine Marie Moacdieh is an Assistant professor in the School of Computer Science at Carleton University. She obtained her Ph.D. in Industrial and Operations Engineering from the University of Michigan, Ann Arbor, in 2015. This work was carried out while she was an Assistant professor in the Department of Industrial Engineering and Management at the American University of Beirut.