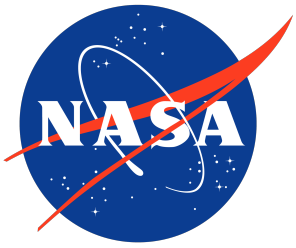
NASA/TM–20250008392



Human Visual Processes for Monitoring Multi-Vehicle Operations

Elliot L. Biltekoff, Tyler D. Fettrow, Michael S. Politowicz, Eric T. Chancey, and Kathryn M. Ballard
NASA Langley Research Center, Hampton, Virginia

December 2025

NASA STI Program Report Series

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA scientific and technical information (STI) program plays a key part in helping NASA maintain this important role.

The NASA STI Program operates under the auspices of the Agency Chief Information Officer. It collects, organizes, provides for archiving, and disseminates NASA's STI. The NASA STI Program provides access to the NTRS Registered and its public interface, the NASA Technical Report Server, thus providing one of the largest collections of aeronautical and space science STI in the world. Results are published in both non-NASA channels and by NASA in the NASA STI Report Series, which includes the following report types:

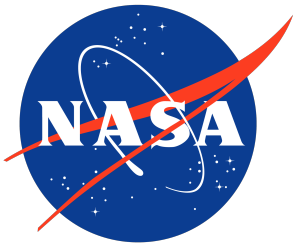
- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA counterpart of peer-reviewed formal professional papers, but having less stringent limitations on manuscript length and extent of graphic presentations.
- TECHNICAL MEMORANDUM. Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.
- CONTRACTOR REPORT. Scientific and technical findings by NASA-sponsored contractors and grantees.
- CONFERENCE PUBLICATION. Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or co-sponsored by NASA.
- SPECIAL PUBLICATION. Scientific, technical, or historical information from NASA programs, projects, and missions, often concerned with subjects having substantial public interest.
- TECHNICAL TRANSLATION. English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services also include organizing and publishing research results, distributing specialized research announcements and feeds, providing information desk and personal search support, and enabling data exchange services.

For more information about the NASA STI Program, see the following:

- Access the NASA STI program home page at `http://www.sti.nasa.gov`
- Help desk contact information: `https://www.sti.nasa.gov/sti-contact-form/` and select the "General" help request type.

NASA/TM–20250008392



Human Visual Processes for Monitoring Multi-Vehicle Operations

Elliot L. Biltekoff, Tyler D. Fettrow, Michael S. Politowicz, Eric T. Chancey, and Kathryn M. Ballard
NASA Langley Research Center, Hampton, Virginia

National Aeronautics and
Space Administration

Langley Research Center
Hampton, Virginia 23681-2199

December 2025

Available from:

NASA STI Program / Mail Stop 150
NASA Langley Research Center
Hampton, VA 23681-2199

Abstract

Mature Urban Air Mobility (UAM) operations will require operators to manage multiple uncrewed and increasingly autonomous aircraft simultaneously. Operators will supervise automated operations rather than directly control individual aircraft, with monitoring as their primary task. Exploring this operational paradigm, we conducted a study to understand the effects of interruptions and set size (i.e., number of vehicles) on operators' situation awareness. In this study, participants performed a multi-vehicle monitoring task while eye movements were recorded. Eye tracking data were filtered and processed, and various metrics were computed to characterize and better understand eye gaze behavior of operators during multi-vehicle monitoring. The present analyses primarily tested set-size effects on eye tracking metrics but also explored the efficacy of using gaze entropy metrics to assess monitoring performance. Results indicated relationships between gaze entropy metrics and set size. Additionally, the findings suggest that the multi-vehicle monitoring task can be treated as a multiple identity tracking (MIT) task rather than a multiple object tracking (MOT) task. This exploratory research provides valuable insights for future research focused on developing passive performance metrics for multi-vehicle monitoring.

Contents

List of Tables

List of Figures

Acronyms

AAM	Advanced Air Mobility
AOI	Area of Interest
eVTOL	Electric Vertical Takeoff and Landing
FW-NSGE	Frequency-Weighted Normalized Stationary Gaze Entropy
IBW	Information Bandwidth
MIT	Multiple Identity Tracking
MOT	Multiple Object Tracking
NGTE	Normalized Gaze Transition Entropy
SAGAT	Situation Awareness Global Assessment Technique
TW-NSGE	Time-Weighted Normalized Stationary Gaze Entropy
UAM	Urban Air Mobility
UAS	Uncrewed Aircraft System

1 Introduction

In some future aviation operations, remote supervisory control of multiple vehicles is expected to become the central responsibility for human operators, with monitoring serving as their main task. The primary purpose of the current study was to examine how interruptions affect an operator's ability to maintain situation awareness during multi-vehicle monitoring. However, the purpose of the current paper is to analyze data from this study to explore the broader potential of using eye tracking to understand monitoring behavior. Specifically, we ultimately seek to develop non-invasive methods for quantifying monitoring performance that can be computed in real-time. As a first step, we examined how participants approached the task in terms of eye gaze behavior to better inform performance metrics. This analysis is a critical step toward developing gaze-based metrics for quantifying and characterizing monitoring performance. Consequently, this paper centers on characterizing and analyzing the eye gaze behavior participants demonstrated during a multi-vehicle monitoring task.

1.1 Operational Context

Advanced Air Mobility (AAM) encompasses various emerging aviation applications, such as Urban Air Mobility (UAM; e.g., air taxis), which aim to transport both passengers and cargo in areas under-served or un-served by current aviation operations (see National Academies of Sciences, Engineering, and Medicine, 2020). Certain AAM applications may adopt an *m:N* operational model, in which a smaller number of human operators (m) remotely oversee a larger number of increasingly autonomous aircraft (N; Aubuchon et al., 2022). The *m:N* paradigm builds on the multi-vehicle operational concept, which posits that a single individual can manage multiple vehicles simultaneously (cf. Lewis, 2013). Mature UAM operations will depend on multi-vehicle (*m:N*) frameworks as a crucial factor for scalability (Patterson et al., 2021). As operations become increasingly autonomous, human interactions with vehicles will shift toward supervisory roles (National Research Council, 2014), with monitoring as the operator's primary task. This emphasizes the pressing need to define performance metrics for monitoring.

1.2 MOT/MIT and Monitoring in Multi-Vehicle (*m:N*) Operations

The literature on multiple object tracking (MOT; Pylyshyn & Storm, 1988) and multiple identity tracking (MIT; Horowitz et al., 2007; Oksama & Hyönä, 2004, 2008) tasks is highly relevant to monitoring in multi-vehicle operations. Although similar, MOT and MIT tasks have key differences. In MOT tasks, observers track a set of visually identical objects (e.g., four identical black circles). In contrast, MIT tasks require observers to track different objects, each with a unique identity. This necessitates binding and updating both identity and location information for the objects. These bindings are maintained through overt, serial attention shifts, as evidenced by gaze location shifts to sample the tracked objects. Evidence suggests that the cognitive systems used for tracking position and identity are distinct (Oksama & Hyönä, 2016). Both MOT and MIT are associated with specific phenomena and effects of task parameters on eye gaze behavior.

The task performed by participants in the current study was not explicitly framed as an MIT task. However, the multi-vehicle monitoring task shared enough similarities

with MIT tasks—such as requiring participants to track both the location and dynamic values of multiple objects moving on a display over time—to infer that the phenomena and effects associated with MIT tasks would emerge. This paper focuses on gathering evidence to explore that possibility in detail. Although previous studies have examined air traffic control-type tasks through the theoretical framework of MIT (e.g., Buck, 2019; Hope, 2009; Hope et al., 2010; Nalbandian & Rantanen, 2015), this study is distinct due to its task complexity and ecological sensitivity in relation to anticipated multi-vehicle UAM operations.

1.3 Eye Tracking

Eye tracking is a well-studied physiological measurement that monitors the position and movement of the eyes, providing insights into visual attention, cognitive processes, and neurological health (Anastasio et al., 2022; Buswell, 1935). In aviation, eye tracking has been used to inform recommendations aimed at improving safety and performance, including analyzing pilots' gaze patterns to detect fatigue, workload, and situation awareness (see Peißl et al., 2018). Similarly, in uncrewed aircraft system (UAS) operations, eye tracking has been widely used to assess operator mental workload (Russo et al., 2025) and also as an innovative control method (Yang et al., 2024). For visual monitoring tasks, gaze data offers significant potential as a non-invasive method for inferring insights about the operator, enabling researchers to explore mechanisms for "closing the loop" between humans and the automation they monitor (Belardinelli, 2024).

Eyes perform saccadic movements to fixate on relevant regions, enabling high-resolution sampling of visual information (Guerrasio et al., 2010). Fixation behavior reflects both internal (top-down) and external (bottom-up) factors, with interactions between perceptual, cognitive, and motor processes (Corbetta & Shulman, 2002; Gordon et al., 2019; Wickens et al., 2022). Derivatives from raw eye tracking data provide insights into internal cognitive and neural processes as well as the external environment (Mahanama et al., 2022). Commonly used metrics include fixation duration, saccadic velocity, and blink frequency. Quantifying area of interest (AOI) metrics, such as the number of times an object is revisited or transitions between objects, provides a broader understanding of gaze behavior (Ponsoda et al., 1995). Additionally, gaze complexity metrics have been useful for analyzing behavioral patterns (Krejtz et al., 2015; Schieber & Gilland, 2008). In this study, we primarily focused on the subset of metrics most relevant to MIT tasks, as defined by prior literature, but we also explored complementary metrics to deepen our characterization.

1.4 Current Study

We conducted an experiment investigating the effects of interruptions and set size (i.e., number of vehicles) on operators' situation awareness (Endsley, 1995b). Situation awareness was operationalized using the Situation Awareness Global Assessment Technique (SAGAT; Endsley, 1995a), which served as the primary performance metric for the monitoring task. The task configuration and associated display were designed to reflect anticipated realistic scenarios that a remote operator might encounter during future UAM operations. Another goal of the experiment, and the focal point of this paper, was to capture and analyze operators' eye gaze behavior during their primary monitoring task. This paper presents an

analysis of the eye tracking data, emphasizing how various eye tracking metrics differ across the independent variables, particularly the number of vehicles being monitored. Metrics were selected based on precedent MIT literature and emerging gaze entropy measures. We summarize the results, discuss their alignment with prior MIT research, and outline future research directions to extend these findings.

2 Method

2.1 Participants

Thirty-five commercial pilots, private pilots, and/or Part 107 certificate holders (i.e., drone pilots) were recruited for the study (33 men, 2 women; mean age = 44.1 years). However, only 18 participants (all men; mean age = 47.0 years) had sufficiently high-quality eye tracking data to be included in the current analyses (see Section 2.3.1). Participants were compensated $350 for participating in the study. This study was approved by NASA's Institutional Review Board (STUDY00000796). Informed consent was obtained from each participant.

2.2 Experimental Tasks

Participants engaged in a simulated multi-vehicle monitoring task, acting as supervisors for electric vertical takeoff and landing (eVTOL) vehicles operating along predefined flight paths across five distinct maps (Dothan, Greensboro, Knoxville, Pittsburgh, and Syracuse). Scenarios involved two or four vehicles traveling between vertiports (i.e., takeoff and landing areas; see Figure 1), and participants were tasked with maintaining situation awareness of vehicle position, operational metrics (e.g., groundspeed, battery percentage, altitude, wind direction, wind speed), and environmental factors (e.g., proximity to noise abatement areas). Vehicle operations were presented as pre-recorded videos of the MPATH (Measuring Performance for Autonomy Teaming with Humans) Ground Control Station (Politowicz et al., 2023) display showing simulated flight operations. Pauses were introduced at specified intervals for SAGAT queries. All participant interactions were performed using a mouse, and the experimental setup was designed to approximate realistic future supervisory control environments. At experimentally derived intervals (see Section 2.4), participants experienced interruptions in the form of a visuo-spatial n-back task, which required them to identify whether the spatial location of a stimulus matched the location of a stimulus presented three trials prior (Hockey & Geffen, 2004). This introduced cognitive demands designed to simulate interruptions that might occur in actual operational environments. Throughout the task, eye tracking data were collected to analyze gaze behavior associated with monitoring performance.

2.3 Eye Tracking

The eye tracking data were acquired using Tobii Pro Spark eye trackers. Data were streamed and written to a file at 60 Hz using custom Python software that was built with the PyGaze library (Dalmaijer et al., 2014). Raw eye gaze coordinates were provided in screen coordinates of a 1080×1920-pixel display and were written to files locally with microsecond

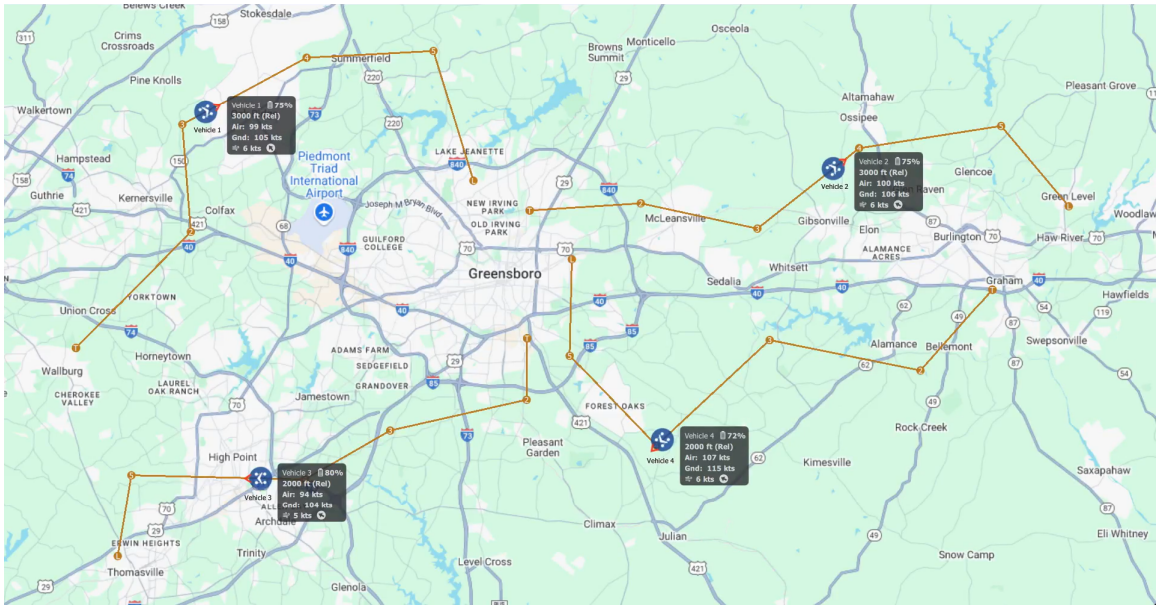


Figure 1. Snapshot of a 4-vehicle scenario (Greensboro). Vehicles are displayed as directional icons on top of their full routes, each with an adjacent text box containing state information. Vehicle parameters, including wind behavior, were continuously updating and displayed.

timestamps. Each eye tracker recording was initiated before each trial and began with a calibration and a validation procedure. Both procedures involved looking at 5 dots (4 corners and the center of screen) sequentially. The validation data were written to file locally. Due to time constraints, we limited the calibration attempts to three per trial. If the calibration was not sufficiently aligned by the third attempt (based on observed results), eye tracking data were still recorded, but poor quality data were filtered out prior to analyses (see Section 2.3.1).

2.3.1 Data Processing

The consistent data log for each scenario in the monitoring task (includes vehicle states, flight path segment locations, and noise abatement area locations) was generated through a combination of two methods: (1) unpacking the raw aircraft messages (MAVLink protocol) from the simulated MPATH flights that were recorded in the scenario video, and (2) performing computer vision analysis with OpenCV (https://opencv.org/) on the scenario video. The monitoring task logs and participant eye tracking logs were then synced for each participant, enabling single data frames with all data of interest for the current analyses. This allowed the research team to determine what participants were looking at throughout the trial and to generate new videos with the eye tracking overlaid, enabling manual verification and quality assessment.

Several AOIs were defined for the monitoring task. The AOIs included the vehicle locations, the location of the text box associated with each vehicle, the trajectory segments, and the noise abatement areas. Hits on AOIs were determined by identifying the AOI that

had the largest surface area within a 50-pixel diameter of the gaze point. This method was used to account for small systematic offsets in the recorded gaze position due to calibration error.

There were varying amounts of noise in the eye tracking data among and within participants from trial to trial. Therefore, we performed smoothing of the eye tracking time series with a 2$^{\text{nd}}$ order Savgol filter with a 5-frame window and then calculated a custom quality score for each participant's eye tracking data for each trial. Because the primary outcome of the eye tracking analysis was AOI hits, we used the AOI hits to define the quality score, Q_{score}, which is derived from the probabilities of various types of AOI hits during the trial. The quality score formula is:

$$Q_{\text{score}} = p_{\text{good}} + p_{\text{bad}} + p_{\text{unknown\_good}} + p_{\text{unknown\_bad}}, \tag{1}$$

where each p term represents a probability associated with AOI hits weighted by their assigned relevance to data quality. These p terms are defined as:

- p_{good}: Probability of "good" AOI hits (fixations on meaningful task elements), positively weighted.
$$p_{\text{good}} = \frac{\sum \text{aoi\_counts[aoi\_good]}}{\text{len(gaze\_data)}} \tag{2}$$

- p_{bad}: Probability of "bad" AOI hits (blinks or off-screen areas), negatively weighted.
$$p_{\text{bad}} = -1 \times \frac{\sum \text{aoi\_counts[aoi\_bad]}}{\text{len(gaze\_data)}} \tag{3}$$

- $p_{\text{unknown\_good}}$: Probability of "unknown good" AOI hits (ambiguous fixations), positively weighted but reduced (0.1).
$$p_{\text{unknown\_good}} = 0.1 \times \frac{\sum \text{aoi\_counts[aoi\_unknown\_good]}}{\text{len(gaze\_data)}} \tag{4}$$

- $p_{\text{unknown\_bad}}$: Probability of "unknown bad" AOI hits (uncertain fixations in irrelevant areas), negatively weighted but reduced (-0.1).
$$p_{\text{unknown\_bad}} = -0.1 \times \frac{\sum \text{aoi\_counts[aoi\_unknown\_bad]}}{\text{len(gaze\_data)}} \tag{5}$$

The weighting reflects the importance of each term, where "good" hits increase the score, "bad" hits decrease it, and "unknown" hits (i.e., ambiguous fixations) have smaller weights. The overall quality score for each trial was calculated by summing these weighted probabilities, with higher values indicating better quality data. Participants that had >20% of trials with quality scores below the chosen threshold were flagged for exclusion from further analysis. The threshold value was chosen based on our manual review of the videos with overlaid gaze to determine a threshold that would appropriately filter subpar data for the current analyses.

2.3.2 Metrics

Several metrics were derived from the raw eye tracking data, most of which have been used in the precedent MIT literature (Hyönä et al., 2019; Li et al., 2018, 2019; Oksama & Hyönä, 2016; Oksama & Hyönä, 2004, 2008). For the primary analyses, each metric was averaged over the trial, excluding the time when there was an n-back interruption or SAGAT probe. Data indices with poor quality (e.g., low capture quality or a blink) were removed. Note that multiple sequential fixations on the same AOI constitute a single target visit; in such cases, the target visit duration is the sum of the individual fixation durations.

Entropy metrics, widely used in eye-tracking analyses (Shiferaw et al., 2019a) outside of MIT literature, were derived from AOI target visit sequences. We used transition matrices, comprised of conditional transition probabilities between AOIs, to compute normalized gaze transition entropy (NGTE). Stationary gaze entropy is typically computed via counts and is referred to here as frequency-weighted normalized stationary gaze entropy (FW-NSGE). We also included an alternative formulation, time-weighted normalized stationary gaze entropy (TW-NSGE), to see if using durations of target visits resulted in a difference. Note that all entropy measures were normalized for two reasons. First, we wanted to be able to compare between set-size conditions (i.e., number of vehicles), which doubled the number of AOIs. Second, normalization of entropy measures facilitates cross-study comparisons (Shiferaw et al., 2019a). All metrics included in the current analysis are defined as follows:

- **Fixation Rate**: Total number of fixations made per second over a trial.

$$\text{Fixation Rate} = \frac{\text{Total Number of All Fixations}}{\text{Total Time (s)}} \tag{6}$$

- **Target Visit Rate**: Total number of times AOIs are visited per second over a trial.

$$\text{Target Visit Rate} = \frac{\text{Total Number of AOI Target Visits}}{\text{Total Time (s)}} \tag{7}$$

- **Mean Fixation Duration**: Average duration of all fixations made over a trial, in seconds.

$$\text{Fixation Duration (s)} = \frac{1}{N}\sum_{i=1}^{N}(\text{End Time}_i - \text{Start Time}_i) \tag{8}$$

  where N is the number of fixations, and End Time_i and Start Time_i are the end and start times of the i-th fixation.

- **Saccadic Velocity**: Velocity of saccadic eye movements, which are rapid movements between fixations.

$$\text{Saccadic Velocity} = \frac{\text{Distance between Fixations}}{\text{Time between Fixations}} \tag{9}$$

- **AOI Fixation Count Proportion**: Proportion of fixations that occur on AOIs compared to the total number of all fixations made.

$$\text{AOI Fixation Count Proportion} = \frac{F_{AOI}}{F_{AOI} + F_{nonAOI}} \tag{10}$$

where F_{AOI} is the fixation count in AOIs, and F_{nonAOI} is the fixation count in non-AOIs.

- **Fixation Duration Ratio (AOI vs Non-AOI)**: Ratio of average AOI fixation durations compared to average non-AOI fixation durations.

$$\text{AOI Fixation Duration Ratio} = \frac{D_{AOI}}{D_{nonAOI}} \tag{11}$$

where D_{AOI} is the average AOI fixation duration, and D_{nonAOI} is the average non-AOI fixation duration.

- **Frequency-Weighted Normalized Stationary Gaze Entropy (FW-NSGE)**: A measure of the uncertainty of the distribution of target visits, computed from total target visit counts.

$$\text{FW-NSGE} = \frac{-\sum_{i \in S} p_i \log_2 p_i}{H_{max}} \tag{12}$$

where S is the set of AOIs, p_i is the probability of a target visit occurring at AOI_i, and H_{max} is equal to log_2(number of AOIs).

- **Time-Weighted Normalized Stationary Gaze Entropy (TW-NSGE)**: Similar to FW-NSGE, but the difference is that probabilities are computed from sums of target visit durations rather than total target visit counts.

$$\text{TW-NSGE} = \frac{-\sum_{i \in S} p_i \log_2 p_i}{H_{max}} \tag{13}$$

where S is the set of AOIs, p_i is the probability of a target visit occurring at AOI_i (as derived from sums of target visit durations), and H_{max} is equal to log_2(number of AOIs).

- **Normalized Gaze Transition Entropy (NGTE)**: A measure of the uncertainty of gaze transitions between AOIs.

$$\text{NGTE} = \frac{-\sum_{i \in S} \pi_i \sum_{j \in S} p_{ij} \log_2 p_{ij}}{H_{max}} \tag{14}$$

where S is the set of AOIs, π_i is the stationary probability of AOI_i, p_{ij} is the transition probability from AOI_i to AOI_j, and H_{max} is equal to log_2(number of AOIs).

We also measured the correlation between the time elapsed since the last visual fixation on a vehicle and the magnitude of error in the reported vehicle location from the SAGAT. The first SAGAT probe required participants to drag and drop icons representing each vehicle to corresponding locations on the map. Location error was measured as the deviation (in pixels) between the actual vehicle location and the dropped icon. This specific analysis examined the relationship between gaze, memory, and vehicle positioning accuracy, which has been explored in the MIT literature (Li et al., 2018).

2.4 Experimental Design

We employed a 2 (interruption rate: 9 or 18 interruptions per trial) $\times$ 2 (interruption length: 14 s or 28 s) $\times$ 2 (set size: 2 vehicles or 4 vehicles) within-subjects design with a control condition that included only the set-size manipulation without any interruptions. The experiment consisted of 10 different scenarios, resulting in a total of 10 trials. Each scenario/trial lasted 630 seconds. Each participant saw the exact same scenario videos. However, the timing of interruptions was pseudo-randomly applied such that an exact set of interruptions, for a given map, was not repeated across participants.

2.5 Procedure

Up to six participants could participate simultaneously in a single session. Upon arrival, participants signed an informed consent form, signed a Privacy Act Notice form, and completed a demographics questionnaire prior to being briefed on the study. The participants then completed approximately 1.5 hours of training, which included watching instructional videos and practicing the monitoring task, SAGAT probes, and interruption task (n-back). Once participants indicated confidence in the protocol, they were offered a break prior to beginning data collection. Each session lasted approximately 9 hours, including a 1-hour lunch break. At the end of the session, participants were debriefed, thanked, and dismissed.

3 Results

We used linear mixed models to assess the significance of the factors of set size (2 vehicles vs. 4 vehicles) and presence of interruptions (control trials vs. interruption trials) on the metrics listed in Section 2.3.2. We assessed trial order as a fixed effect to confirm that the long study session was not inducing fatigue as a confound. The effect of trial order was not significant across all metrics, so we removed it from the model. Additionally, we assessed scenario as a fixed effect in case particular map layouts with vehicle trajectories were introducing a confound. However, the effect of scenario was also not significant across all metrics, so we removed it from the model. Participants were treated as a random effect to account for their repeated measures. For each model, Bayes Factors (BF) were computed by comparing the Bayesian Information Criterion (BIC) of the full model to the BIC of reduced models, each with an excluded individual fixed effect (Raftery, 1995).

We confidently rejected the null hypothesis of a factor's non-significance (i.e., we rejected a factor's contribution as irrelevant) if both of the following were true:

1. The 95% confidence intervals of the parameter estimate for a factor with $p < 0.05$ excluded zero; and

2. The Bayes Factor for that factor was greater than 1 (BF > 1), constituting evidence for the full model over the reduced model.

For BF > 1 (i.e., evidence for the full model over the reduced model), we adopted the following qualitative interpretation for strength of evidence from Jeffreys (1961; see Wetzels et al., 2011): BF $1-3$ constitutes *anecdotal* evidence, BF $3-10$ constitutes *substantial* evidence, BF $10-30$ constitutes *strong* evidence, BF $30-100$ constitutes *very strong* evidence, and BF > 100 constitutes *decisive* evidence. As Bayes Factors grow extraordinarily high (as was the result for several effects in this study), the likelihood that the reduced model is superior to the full model becomes infinitesimally small. Note that BF < 1 constitutes evidence for the reduced model over the full model.

The primary task involved maintaining situation awareness of 2 or 4 vehicles, depending on the scenario. This task required constant scanning of each vehicle's position and heading (vehicle icon), state information (adjacent text box), as well as the projected flight path and noise abatement areas. Figure 2 displays the fixation time spent on each AOI as a percentage of total fixation time for the 2- and 4-vehicle conditions.

The statistical results from the metric analyses mixed models and BF calculations are provided in Table 1. Every metric analyzed had a significant effect of set size (number of vehicles) except saccadic velocity. Fixation rate, target visit rate, and mean fixation duration were the only three metrics to show a significant effect of interruption with a BF > 1. Fixation rate and target visit rate increased when there were interruptions, whereas the mean fixation duration decreased.

Figure 3 provides a visualization of the results for set size, as this was the primary term of interest (and had BF > 500 in most instances). The plots depict the individual data points (color coded by participant) encased by a violin plot to provide distribution context. The white line that extends across the violin plot is the calculated mean of the distribution.

Fixation rate and target visit rate increased from 2 to 4 vehicles, whereas the mean fixation duration decreased from 2 to 4 vehicles. Counts and mean durations of fixations made on AOIs, relative to fixations on non-AOIs, increased from 2 to 4 vehicles. Neither factor had a significant effect on saccadic velocity. Both the FW-NSGE and TW-NSGE increased substantially from 2 to 4 vehicles. The NGTE decreased from 2 to 4 vehicles.

The relationship between time since last fixation on a vehicle icon and accuracy of vehicle location identification was significantly positive, $r = .11$, $p < .001$ (see Figure 4). The weak positive slope suggests that the location error increased slightly as more time elapsed since the participant last viewed the vehicle icon AOI.

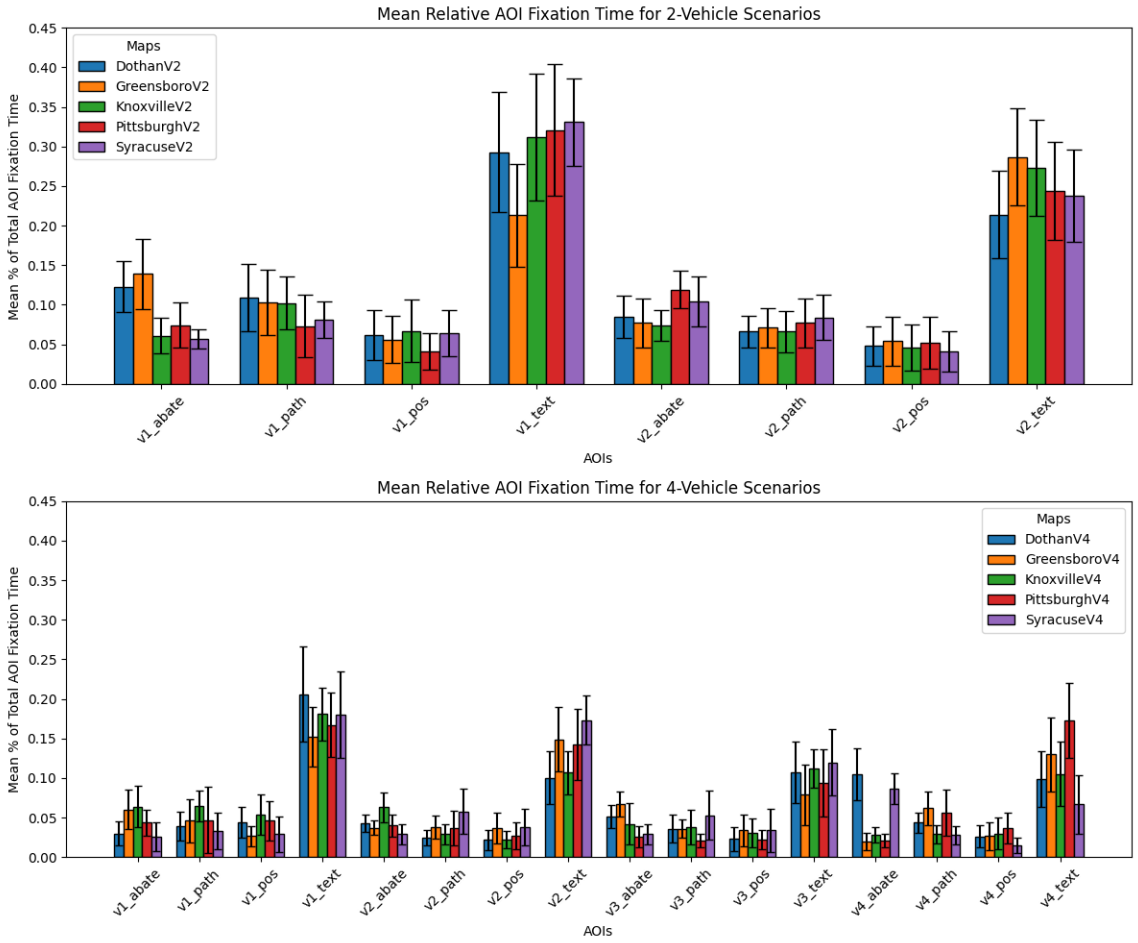


Figure 2. Percent fixation time spent on AOIs for the 2- and 4-vehicle conditions. Values are calculated as a percentage of total fixation time. Error bars represent standard deviations. V1 = Vehicle 1; V2 = Vehicle 2; V3 = Vehicle 3; V4 = Vehicle 4; abatement/abate = the noise abatement area associated with a specific vehicle's flight path; path = the specific vehicle's indicated flight path drawn on the map; position/pos = the drone icon associated with a specific vehicle; text = the black data tag associated with a specific vehicle. (See Figure 1 for depiction of AOIs.)

Table 1. Statistical results for eye tracking metrics.

	Bayes Factor	Coef.	Std. Error	*z*	*p*	[0.025	0.975]
Fixation Rate							
Intercept	NaN	3.299	0.076	43.296	0.000	3.150	3.449
set-size_4[T.True]	**552.133**	**0.159**	**0.037**	**4.343**	**0.000**	**0.087**	**0.231**
interruption[T.True]	**1.115**	**0.107**	**0.046**	**2.342**	**0.019**	**0.018**	**0.197**
Target Visit Rate							
Intercept	NaN	1.635	0.071	22.925	0.000	1.495	1.774
set-size_4[T.True]	**7.37E+25**	**0.499**	**0.036**	**13.744**	**0.000**	**0.428**	**0.570**
interruption[T.True]	**1250.890**	**0.206**	**0.045**	**4.547**	**0.000**	**0.117**	**0.296**
Mean Fixation Duration							
Intercept	NaN	0.273	0.006	45.099	0.000	0.261	0.284
set-size_4[T.True]	**6307.390**	**-0.015**	**0.003**	**-4.940**	**0.000**	**-0.021**	**-0.009**
interruption[T.True]	**2.111**	**-0.010**	**0.004**	**-2.610**	**0.009**	**-0.018**	**-0.003**
Saccadic Velocity							
Intercept	NaN	56.547	2.191	25.812	0.000	52.253	60.841
set-size_4[T.True]	0.237	-2.232	1.467	-1.522	0.128	-5.108	0.643
interruption[T.True]	0.477	-3.550	1.836	-1.934	0.053	-7.149	0.048
AOI Fixation Count Proportion							
Intercept	NaN	0.863	0.012	74.750	0.000	0.840	0.886
set-size_4[T.True]	**2.68E+06**	**0.041**	**0.007**	**6.227**	**0.000**	**0.028**	**0.054**
interruption[T.True]	0.107	-0.007	0.008	-0.843	0.399	-0.023	0.009
Fixation Duration Ratio (AOI vs Non-AOI)							
Intercept	NaN	1.332	0.060	22.049	0.000	1.214	1.451
set-size_4[T.True]	**8.698**	**0.118**	**0.038**	**3.127**	**0.002**	**0.044**	**0.192**
interruption[T.True]	0.707	-0.100	0.047	-2.131	0.033	-0.192	-0.008
Frequency-Weighted Normalized Stationary Gaze Entropy (FW-NSGE)							
Intercept	NaN	0.714	0.004	174.531	0.000	0.706	0.722
set-size_4[T.True]	**1.11E+149**	**0.230**	**0.002**	**102.532**	**0.000**	**0.225**	**0.234**
interruption[T.True]	0.346	0.005	0.003	1.756	0.079	-0.001	0.010
Time-Weighted Normalized Stationary Gaze Entropy (TW-NSGE)							
Intercept	NaN	0.638	0.011	59.708	0.000	0.617	0.659
set-size_4[T.True]	**2.56E+95**	**0.238**	**0.005**	**49.036**	**0.000**	**0.228**	**0.247**
interruption[T.True]	0.720	0.013	0.006	2.141	0.032	0.001	0.025
Normalized Gaze Transition Entropy (NGTE)							
Intercept	NaN	0.728	0.009	85.225	0.000	0.712	0.745
set-size_4[T.True]	**1.50E+10**	**-0.046**	**0.006**	**-7.828**	**0.000**	**-0.057**	**-0.034**
interruption[T.True]	0.081	0.003	0.007	0.388	0.698	-0.011	0.017

Note. Linear mixed models assessed the effects of set size and interruption on each metric. Rows are bold where Bayes Factor > 1 and 95% confidence intervals exclude zero. Fixed effect terms are dummy coded such that *term*[T.True] is the "treatment group." The coefficient for this term corresponds to the expected change in the metric in question, where the intercept coefficient is effectively the expected measured value of the metric in the reference condition (in this case, a set size of 2 vehicles and no interruptions).

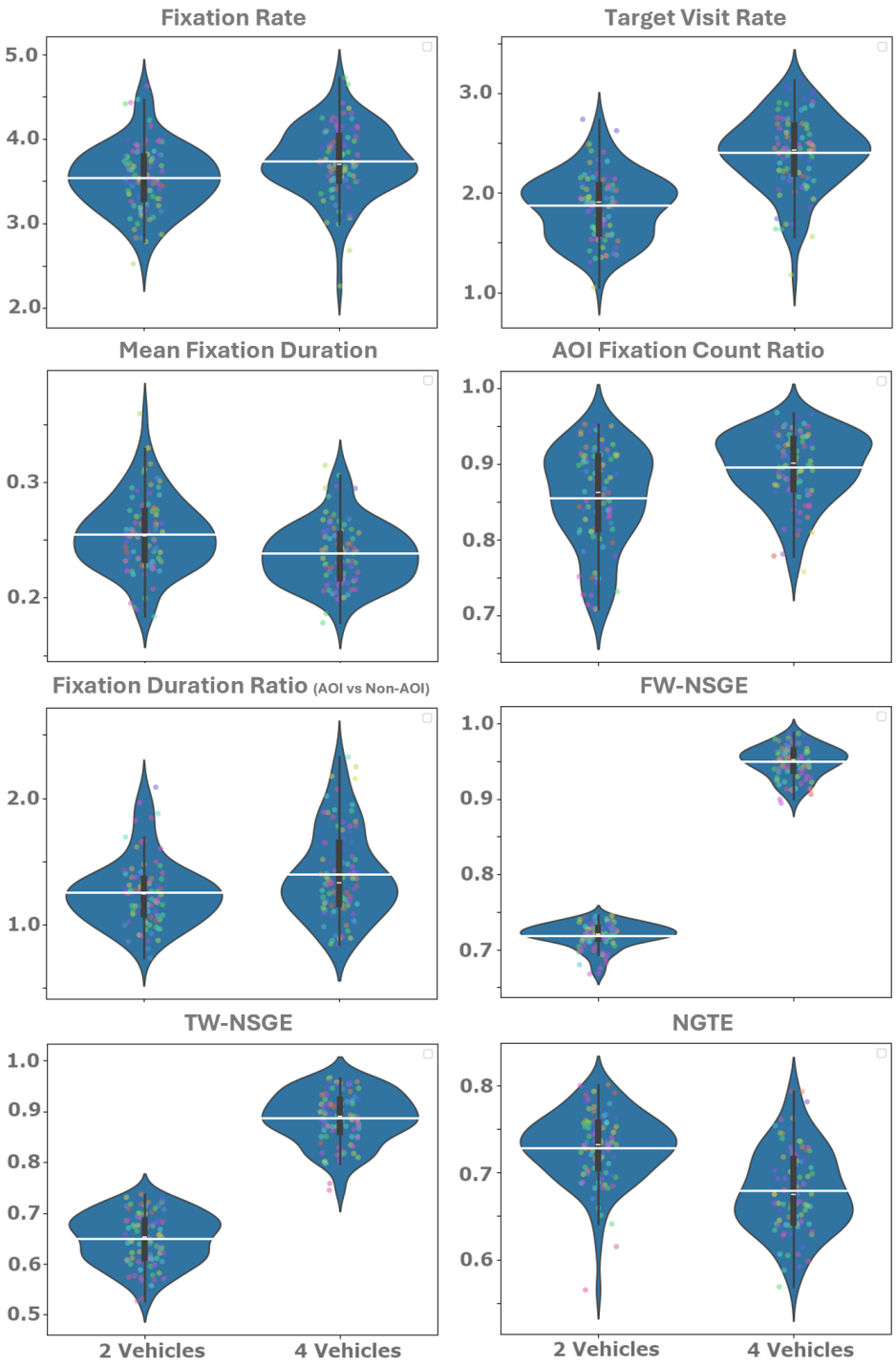


Figure 3. Eye tracking metric results. Violin plots are shown for all eye tracking metrics except saccadic velocity, which was not statistically significant (see Section 2.3.2). For each plot, horizontal white line depicts the mean, dots depict individual trials, and colors depict participants. The x-axis labels apply vertically to all plots.

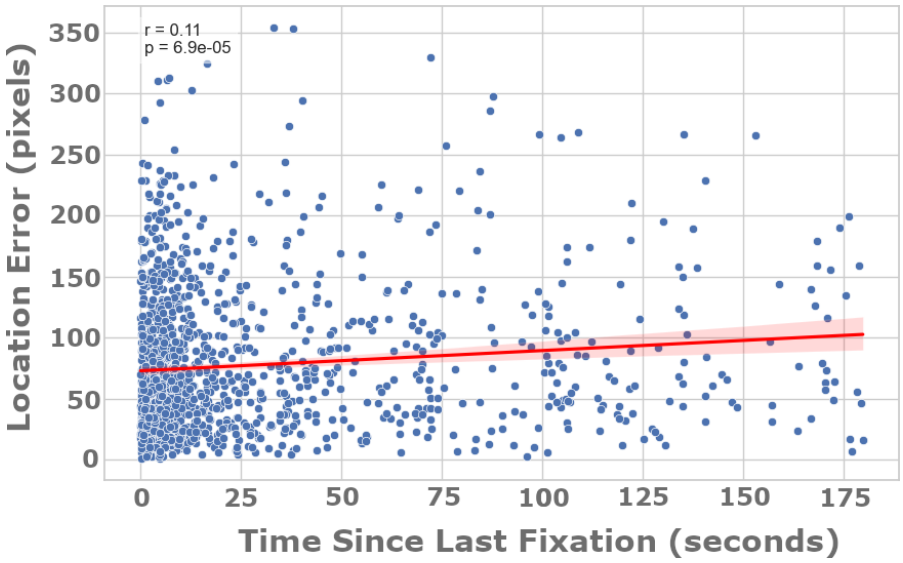


Figure 4. Correlation between last fixation and location error. This plot depicts the relationship between time since last fixation on a vehicle icon and accuracy of vehicle location identification. The red line is the linear regression between the two variables, with r and p displayed in the top left of the figure. The red banding corresponds to the 95% confidence interval.

4 Discussion

4.1 Metrics with Known Relevance to Multiple Identity Tracking

Based on our results, the multi-vehicle monitoring task was likely treated as an MIT task (rather than an MOT task) by the participants, which is reflected in the eye tracking metrics (see Table 2). Both fixation rates and target visit rates increased with set size, consistent with findings from Oksama and Hyönä (2008). This suggests that operators increase their sample rate as the number of AOIs they need to sample grows. Consistent with Oksama and Hyönä (2016), mean fixation duration decreased with a larger set size, suggesting that more frequent sampling is necessary as the number of AOIs increases. The fixation duration ratio (AOI vs non-AOI) metric results aligned with previous findings (Li et al., 2018; Oksama & Hyönä, 2016) showing that operators spent more time and made more frequent fixations on AOIs compared to non-AOIs, with this effect becoming more pronounced as set size increased.

The only analysis that did not support treating the monitoring task as an MIT task, based on precedent literature, was the correlation between the recency of target fixation and location error. The relationship was positive, as expected, but its strength was notably weaker compared to previous findings (Li et al., 2018). A plausible explanation for this result is that the constant visibility of the complete flight path likely accounts for much of the observed correlation. In studies where this correlation was stronger, there was no visible flight path. Additionally, the speed of all vehicle icons across scenarios was relatively

Table 2. Summary of eye tracking metrics, results, and MIT support.

Metric	Result	Supported MIT?
Fixation Rate	Higher set-size increased fixation rate.	✓
Target Visit Rate	Set-size increased target visit rate.	✓
Mean Fixation Duration	Higher set-size reduced fixation durations.	✓
AOI Fixation Count Proportion	More fixations on AOIs than non-AOIs.	✓
Fixation Duration Ratio (AOI vs Non-AOI)	AOI fixations were longer than non-AOI.	✓
Recency of Target Fixation / Tracking Accuracy	Small correlation between target fixation and accuracy.	×
Saccadic Velocity	Set-size and condition affected MSV.	?
NGTE	Higher set-size decreased NGTE.	?
FW-NSGE and TW-NSGE	Higher set-size increased FW-NSGE and TW-NSGE.	?

Note. The results highlight the significant findings in support of the multi-vehicle monitoring task being treated as an MIT task. The green check marks suggest positive support; the red "x" suggests insufficient support; and the "?" suggests that support of MIT is unclear due to insufficient body of work.

slow and consistent, such that vehicle icon locations changed very little even over extended periods. Displacement per second, a metric shown to predict location error (Hope, 2009), was minimal in this context. Therefore, operators could reasonably extrapolate a vehicle's location based on its constant speed and the known flight path, even without having recently fixated on the vehicle.

Gaze behavior metrics derived from the eye tracking data provide sufficient evidence to support treating the multi-vehicle (m:N) monitoring performed by participants as an MIT task. Thus, it is reasonable to expect the findings to translate more broadly to other multi-vehicle monitoring tasks. If these findings do generalize, the MIT literature can be used to guide the design of task configurations, processes, and corresponding displays. For example, evidence suggests that fixation duration and fixation rate have lower bounds relative to set size (Oksama & Hyönä, 2016), which may impact monitoring tasks by potentially creating bottlenecks in performance. This effect appears to result from set size increasing the demands of an MIT task, such that the trade-off between sampling rate and observation time that observers may employ can no longer be exploited, leading to a precipitous decline in tracking accuracy (Moray, 1984).

Interestingly, the effect of interruption increased both fixation and target visit rates, and decreased mean fixation duration. This pattern suggests that interruptions prompted participants to better exploit the trade-off between sampling rate and observation. Specifically, in trials where interruptions occurred, participants appeared to 'make better use of

their time' by sampling more frequently and for shorter durations during the periods when they were on task. However, the effect of interruptions has not been explored in prior MIT literature. Therefore, this finding does not provide additional support beyond the observed set-size effects for treating the task as an MIT task.

4.2 Exploratory Metrics

In addition to the metrics discussed in Section 4.1, we examined the lower-level metric of saccadic velocity, as well as the higher-order metrics of NGTE, TW-NSGE, and FW-NSGE. Table 2 includes question marks next to these metrics because no prior literature provides an a priori basis for hypotheses regarding their relevance to MIT. We found no significant effects of set size or interruptions on saccadic velocity. However, all of the entropy metrics showed significant effects, offering intriguing insights.

We considered the possibility that saccadic velocity may be higher for the 4-vehicle condition given that more fixations are being made, with less time being spent on each fixation. A plausible (though simplistic) hypothesis is that individuals might attempt to 'save time' by shifting their gaze between targets at a faster speed, just as they 'save time' by sampling AOIs more frequently and for shorter durations. However, saccadic velocity is not easily subject to conscious control (Zuber et al., 1965), especially when compared to factors like fixation frequency or duration. Notably, peak saccadic velocity has been shown to be a far more sensitive metric than mean saccadic velocity for detecting changes in task complexity (Di Stasi et al., 2013). Interestingly, there is some evidence that peak saccadic velocity can be slightly influenced voluntarily (Muhammed et al., 2020), unlike mean saccadic velocity. However, due to the sampling rate of the eye-tracking data, we were unable to reliably compute peak saccadic velocity and therefore did not include this metric. Additionally, given the relationship between saccadic velocity and saccadic amplitude (Bahill et al., 1975; Baloh et al., 1975), any decrease in saccadic velocity observed when transitioning from 2- to 4-vehicle scenarios would likely be attributable to a reduction in the average distance between AOIs. In the 4-vehicle scenarios, the number of AOIs doubles but occupies the same amount of space. This idea suggests that if the distance traveled by saccades is generally shorter in one condition compared to another, the mean saccadic velocity is also likely to be slower.

Set size (i.e., number of vehicles) significantly affected all entropy metrics. The effect of set size on NGTE indicates that gaze transitions become more predictable and structured in the higher set-size condition (4 vehicles), likely due to the increased demand of monitoring more AOIs. These findings suggest that, as task demands increase, operators may adopt more systematic gaze patterns to process information more efficiently, as evidenced by the lower NGTE in the 4-vehicle condition. This observation aligns with recent research highlighting gaze transition entropy as a critical metric in dynamic, automation-driven workspaces (Cui et al., 2024).

There was a notable increase in TW-NSGE and FW-NSGE from the 2-vehicle condition to the 4-vehicle condition. This suggests that operator gaze behavior became more uniform as the number of AOIs increased. Di Stasi et al. (2016) provided evidence for an effect of task complexity on FW-NSGE, but the study manipulated task complexity by varying the surgical tasks performed. Thus, comparing those results to the current work is challenging, as task complexity in Di Stasi et al. (2016) was derived by entirely changing the tasks.

Regardless, our findings emphasize the need for further exploration of entropy metrics in relation to task complexity. This brings to light an important consideration for future work: determining how task complexity should be controlled to assess its impact on gaze entropy metrics. Another essential aspect to consider is whether task complexity stems from completely changing the task itself (e.g., different goals or tools available) or from modifying specific aspects of the task (e.g., the number of elements to monitor, the information bandwidth of the display, the level of precision required). Ultimately, understanding the role of the source of task complexity in shaping its effect on entropy metrics remains an open line of inquiry.

Although not explicitly analyzed, FW-NSGE appears higher than TW-NSGE. This suggests that although operators may visit certain AOIs more frequently than others, the disparity becomes more pronounced when viewed through the lens of duration. Specifically, certain AOIs—such as vehicle text boxes—exhibited a disproportionately larger sum of target visit durations on average (see Figure 2) compared to other AOIs, which supports this interpretation.

Finally, operators may prioritize certain AOIs containing more critical information (such as the vehicle text boxes) over others, resulting in lower TW-NSGE compared to FW-NSGE. This prioritization could be influenced by factors such as the task, training, and operator expectations. Related work has explored this concept, such as research on the SEEV (Salience, Effort, Expectancy, Value) model (Wickens et al., 2001; see Wickens et al., 2022).

4.3 Limitations and Considerations

This work provides valuable insights into gaze behavior in the context of monitoring multi-vehicle operations, but it does not come without limitations and considerations. Foremost, it is important to reiterate that eye tracking served as a secondary, exploratory component of the study. Specifically, the experiment was not designed around any particular eye tracking hypotheses, which limited the analyses that could be performed (e.g., time-series analysis; see Section 4.4). Additionally, the eye tracking sampling rate (60 Hz) and data quality (excluded data from 17 participants) also posed limitations. The sampling rate is sufficient for basic eye tracking metrics, but it may not be sufficient for capturing higher frequency gaze patterns like peak saccadic velocity or derivatives of pupillometry data. Future studies could benefit from higher frequency sampling to make use of these more sensitive measures. The loss of subject data due to poor quality most likely stems from individual differences, as well as incompatibility between participant corrective vision prescriptions and the eye tracking sensor. This poor data quality led to a reduction in the available sample size, which in turn affected the statistical power of the analyses. Although significant results were still observed, it is possible that the exclusion of participants diminished the ability to detect additional effects, especially concerning the interaction between interruption rates and set-size conditions. Future studies should consider ensuring compatibility between participants' vision prescriptions and eye tracking equipment to prevent such exclusions. Future work will benefit from addressing these limitations.

4.4 Future Work

Although the current study provides valuable insights into gaze behavior during the monitoring of multi-vehicle operations, further research is needed to refine our understanding of eye tracking metrics and their implications for complex monitoring tasks. Entropy metrics, in particular, show strong potential for revealing how operators manage and respond to informational demands in monitoring tasks through gaze behavior. These metrics should be examined both independently and together, as Shiferaw et al. (2019a) suggest: "Because they measure different aspects of visual scanning, concurrent evaluation of gaze transition entropy and stationary gaze entropy measures may provide a more complete assessment of gaze behavior" (p. 358).

Limited research has examined operators' gaze behavior in relation to informational demands within the context of MIT. However, evidence suggests that operators may, perhaps unconsciously, attempt to "optimally sample" display elements based on informational demands (Oksama & Hyönä, 2016). However, this work was limited to location-identity bindings of targets only, and informational demand was operationalized as the frequency with which objects simultaneously changed both speed and direction. A compelling argument posits that displacement over time may be a better way to operationalize informational demand. However, the study supporting this approach notes that the objects being tracked were moving faster—measured in visual angle per second—than what is typical in most multi-vehicle monitoring tasks (Hope, 2009). Thus, we propose that future research on eye tracking during MIT tasks should consider using displacement over time, rather than speed or direction change frequency, in combination with low object speeds. This approach would enhance the ecological relevance of the study to multi-vehicle monitoring tasks and allow for more precise modulation of informational demand.

The potential relationship between task dynamics and entropy metrics over time merits further exploration. Specifically, we recognized that these metrics lend themselves to being structured as a time series rather than a single value that characterizes an entire scenario. A comprehensive review of entropy metrics and their applications supports this perspective: "The spatial distribution of information changes over time in fully dynamic stimuli. In such cases, calculating entropy rate per a given time bin (depending on the rate of change in the stimuli) may better account for associated changes in visual scanning" (Shiferaw et al., 2019a, p. 363). This approach allows for more nuanced examination of gaze behavior over time, particularly when considering the dynamism of monitoring moving vehicles.

A key step in this direction is quantification of the information bandwidth (IBW), a measure of informational demand (or output) of display elements, as a proxy metric for the overall dynamism of a scenario. This concept is similar to the *value* and *expectancy* parameters in the SEEV model (Wickens et al., 2022), in which the value of the information being presented and the expectation of new information appearing both influence visual attention allocation. We developed an exploratory calculation of IBW that checked for significant changes in various system parameters (e.g., battery percentage, altitude, airspeed, groundspeed, heading) over time. This approach used thresholds for each parameter to track cumulative changes, and when the cumulative change exceeded a set threshold, a "flag" was triggered marking a notable change. This tracking was done for each system parameter, and the overall IBW value was the sum of the individual system parameter IBW values. Figure 5 displays a time series representation of IBW along with gaze entropy

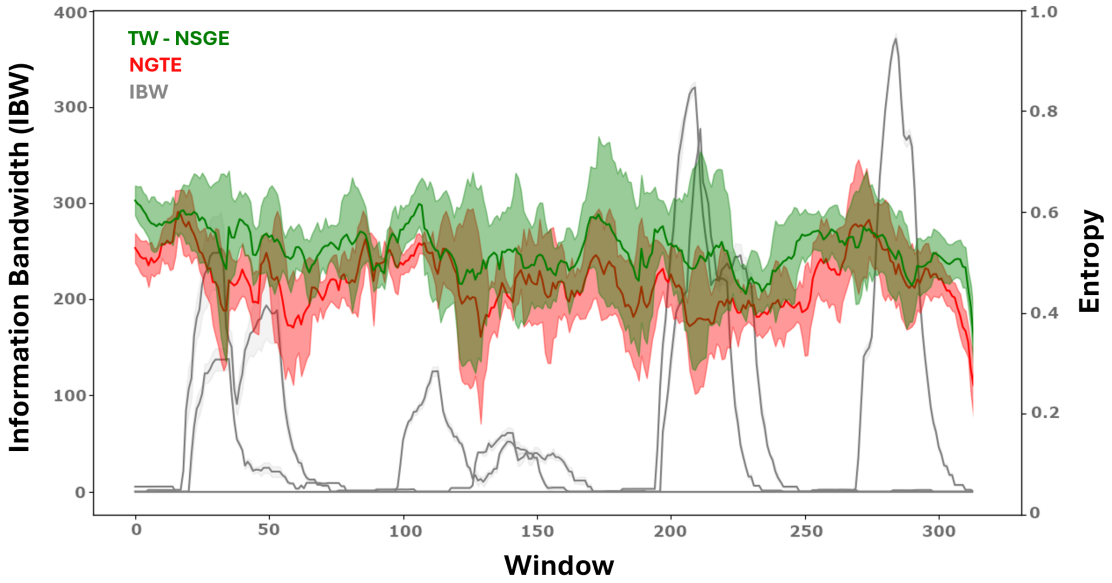


Figure 5. Time series data for information bandwidth and entropy. Overall IBW (for each vehicle in a 2-vehicle scenario), TW-NSGE, and NGTE are displayed over time. The gaze entropy lines depict data from all participants during the same 2-vehicle scenario. The x-axis is in units of time windows that were used for calculating the variables (30-second windows with 2-second steps). Bands correspond to 95% confidence intervals.

metrics (TW-NSGE and NGTE). The figure provides insight into potential patterns and relationships that are not immediately apparent in static statistical models. However, due to the absence of control over some experimental variables (e.g., IBW) and the applied nature of the current study, further work is needed to effectively establish relationships over time. Additionally, further refinement of the IBW calculation is necessary to more accurately reflect the relative magnitude of an operator's perceived value and expectation of specific information on the display.

Ultimately, a deeper understanding of typical gaze behavior is necessary for envisioned multi-vehicle monitoring tasks, particularly with varying task parameters and atypical operator states. Additionally, the potential of using entropy metrics (i.e., NGTE, FW-NSGE, and TW-NSGE) as robust metrics for capturing gaze behavior warrants further exploration. Future research should pursue a deeper examination of the meaning and applicability of these metrics as tools for measuring performance during monitoring tasks. Given previous literature demonstrating the potential efficacy of stationary gaze entropy for inferring change detection (Radun et al., 2017), entropy metrics could potentially be applied to tasks where monitoring is a prerequisite, such as detection or supervisory control tasks. Furthermore, entropy metrics demonstrate sensitivity to physiological state changes, such as the consumption of exogenous substances like alcohol (Shiferaw et al., 2019b) and the onset of fatigue (Naeeri et al., 2022), which could significantly impact performance in monitoring or related tasks.

By addressing the limitations of the current study and refining our approach in future studies, we aim to provide a more robust understanding of how gaze metrics like entropy can inform the design of systems that support multi-vehicle monitoring tasks with safety and efficiency in mind.

5 Conclusion

We conducted a study in which commercial pilots, private pilots, and Part 107 certificate holders (i.e., drone pilots) performed a multi-vehicle monitoring task. Eye tracking data were collected during the experiment and then filtered, processed, and used to compute various metrics. These metrics included lower-level metrics such as fixations and target visits per second, mean fixation duration, AOI fixation count proportion, and fixation duration ratio (AOI vs non-AOI). The metrics were compared to findings from precedent MIT literature to evaluate whether multi-vehicle monitoring tasks could be considered analogous to MIT tasks.

We argue that our results provide sufficient evidence to support this claim, as all observed effects (except for a single analysis that fell outside the primary focus of the study) aligned with findings in precedent MIT literature. We also conducted several additional analyses using metrics that have not been explicitly examined in the context of MIT tasks, including higher-level gaze metrics such as NGTE and FW-NSGE/TW-NSGE. The purpose of these analyses was to identify potential set-size or interruption effects on these metrics. Set-size effects were observed for all gaze entropy metrics. We connected our findings to precedent literature on the various applications of gaze entropy metrics and identified a promising opportunity to use these metrics as real-time performance measures for multi-vehicle monitoring tasks, potentially in combination with some of the lower-level metrics analyzed in this study. In doing so, we have provided insights into expected gaze behavior in MIT tasks, particularly concerning the task parameter of set size, which can be leveraged within the multi-vehicle operational space.

We offered a basic example of a promising future research avenue, along with additional suggestions for potential directions. Ultimately, this work has made considerable progress toward using gaze behavior metrics derived from eye tracking data as passive performance measures for multi-vehicle monitoring.

References

Anastasio, T. J., Demer, J. L., Leigh, R. J., Luebke, A. E., van Opstal, A. J., Optican, L. M., Ramat, S., & Zee, D. (Eds.). (2022). *David A. Robinson's modeling the oculomotor control system* (Vol. 267). Elsevier.

Aubuchon, V. V., Hashemi, K. E., Shively, R. J., & Wishart, J. M. (2022). Multi-vehicle (m: N) operations in the NAS - NASA's research plans. *AIAA Aviation 2022 Forum*. https://doi.org/10.2514/6.2022-3758

Bahill, A. T., Clark, M. R., & Stark, L. (1975). The main sequence, a tool for studying human eye movements. *Mathematical Biosciences*, *24*(3-4), 191–204. https://doi.org/10.1016/0025-5564(75)90075-9

Baloh, R. W., Sills, A. W., Kumley, W. E., & Honrubia, V. (1975). Quantitative measurement of saccade amplitude, duration, and velocity. *Neurology*, *25*(11), 1065. https://doi.org/10.1212/WNL.25.11.1065

Belardinelli, A. (2024). Gaze-based intention estimation: Principles, methodologies, and applications in HRI. *ACM Transactions on Human-Robot Interaction*, *13*(3). https://doi.org/10.1145/3656376

Buck, A. (2019). *Multiple identity tracking and motion extrapolation* [Master's thesis, Rochester Institute of Technology].

Buswell, G. T. (1935). *How people look at pictures: A study of the psychology and perception in art*. University of Chicago Press.

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, *3*(3), 201–215. https://doi.org/10.1038/nrn755

Cui, Z., Sato, T., Jackson, A., Jayarathna, S., Itoh, M., & Yamani, Y. (2024). Gaze transition entropy as a measure of attention allocation in a dynamic workspace involving automation. *Scientific Reports*, *14*. https://doi.org/10.1038/s41598-024-74244-4

Dalmaijer, E. S., Mathôt, S., & Van der Stigchel, S. (2014). Pygaze: An open-source, cross-platform toolbox for minimal-effort programming of eyetracking experiments. *Behavior Research Methods*, *46*(4), 913–921. https://doi.org/10.3758/s13428-013-0422-2

Di Stasi, L. L., Diaz-Piedra, C., Rieiro, H., Sanchez Carrion, J. M., Martin Berrido, M., Olivares, G., & Catena, A. (2016). Gaze entropy reflects surgical task load. *Surgical Endoscopy*, *30*, 5034–5043. https://doi.org/10.1007/s00464-016-4851-8

Di Stasi, L., Marchitto, M., Antolí, A., & Cañas, J. (2013). Saccadic peak velocity as an alternative index of operator attention: A short review. *European Review of Applied Psychology*, *63*(6), 335–343. https://doi.org/10.1016/j.erap.2013.09.001

Endsley, M. R. (1995a). Measurement of situation awareness in dynamic systems. *Human Factors*, *37*(1), 65–84. https://doi.org/10.1518/001872095779049499

Endsley, M. R. (1995b). Toward a theory of situation awareness in dynamic systems. *Human Factors*, *37*(1), 32–64. https://doi.org/10.1518/001872095779049543

Gordon, N., Tsuchiya, N., Koenig-Robert, R., & Hohwy, J. (2019). Expectation and attention increase the integration of top-down and bottom-up signals in perception through different pathways. *PLOS Biology*, *17*(4). https://doi.org/10.1371/journal.pbio.3000233

Guerrasio, L., Quinet, J., Büttner, U., & Goffart, L. (2010). Fastigial oculomotor region and the control of foveation during fixation. *Journal of Neurophysiology*, *103*(4), 1988–2001. https://doi.org/10.1152/jn.00771.2009

Hockey, A., & Geffen, G. (2004). The concurrent validity and test–retest reliability of a visuospatial working memory task. *Intelligence*, *32*(6), 591–605. https://doi.org/10.1016/j.intell.2004.07.009

Hope, R. M. (2009). *The predictive utility of the model of multiple identity tracking in air traffic control performance* [Master's thesis, Rochester Institute of Technology].

Hope, R. M., Rantanen, E. M., & Oksama, L. (2010). Multiple identity tracking and entropy in an ATC-like task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *54*(13), 1012–1016. https://doi.org/10.1177/154193121005401303

Horowitz, T. S., Klieger, S. B., Fencsik, D. E., Yang, K. K., Alvarez, G. A., & Wolfe, J. M. (2007). Tracking unique objects. *Perception & Psychophysics*, *69*, 172–184. https://doi.org/10.3758/BF03193740

Hyönä, J., Li, J., & Oksama, L. (2019). Eye behavior during multiple object tracking and multiple identity tracking. *Vision*, *3*(3), 37. https://doi.org/10.3390/vision3030037

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford University Press.

Krejtz, K., Duchowski, A., Szmidt, T., Krejtz, I., González Perilli, F., Pires, A., et al. (2015). Gaze transition entropy. *ACM Transactions on Applied Perception*, *13*, 1–20. https://doi.org/10.1145/2834121

Lewis, M. (2013). Human interaction with multiple remote robots. *Reviews of Human Factors and Ergonomics*, *9*(1), 131–174. https://doi.org/10.1177/1557234X13506688

Li, J., Oksama, L., & Hyönä, J. (2018). Close coupling between eye movements and serial attentional refreshing during multiple-identity tracking. *Journal of Cognitive Psychology*, *30*(5-6), 609–626. https://doi.org/10.1080/20445911.2018.1476517

Li, J., Oksama, L., & Hyönä, J. (2019). Model of Multiple Identity Tracking (MOMIT) 2.0: Resolving the serial vs. parallel controversy in tracking. *Cognition*, *182*, 260–274. https://doi.org/10.1016/j.cognition.2018.10.016

Mahanama, B., Jayawardana, Y., Rengarajan, S., Jayarathna, G., Chukoskie, L., Snider, J., & Jayarathna, S. (2022). Eye movement and pupil measures: A review. *Frontiers in Computer Science*, *3*. https://doi.org/10.3389/fcomp.2021.733531

Moray, N. (1984). Attention to dynamic visual displays in man-machine systems. In R. Parasuraman & D. Davies (Eds.), *Varieties of attention* (pp. 485–513). Academic Press.

Muhammed, K., Dalmaijer, E., Manohar, S., & Husain, M. (2020). Voluntary modulation of saccadic peak velocity associated with individual differences in motivation. *Cortex*, *122*, 198–212. https://doi.org/10.1016/j.cortex.2018.12.001

Naeeri, S. M., Kang, Z., & Palma Fraga, R. (2022). Investigation of pilots' visual entropy and eye fixations for simulated flights consisted of multiple take-offs and landings. *Journal of Aviation/Aerospace Education & Research*, *31*(2). https://doi.org/10.15394/jaaer.2022.1920

Nalbandian, A., & Rantanen, E. (2015). Learning of location-identity bindings: Development of level 1 situation awareness in an air traffic control-like task. *18th International Symposium on Aviation Psychology*, 159–164.

National Academies of Sciences, Engineering, and Medicine. (2020). *Advancing aerial mobility: A national blueprint*. The National Academies Press. https://doi.org/10.17226/25646

National Research Council. (2014). *Autonomy research for civil aviation: Toward a new era of flight*. The National Academies Press. https://doi.org/10.17226/18815

Oksama, L., & Hyönä, J. (2016). Position tracking and identity tracking are separate systems: Evidence from eye movements. *Cognition*, *146*, 393–409. https://doi.org/10.1016/j.cognition.2015.10.016

Oksama, L., & Hyönä, J. (2004). Is multiple object tracking carried out automatically by an early vision mechanism independent of higher-order cognition? An individual difference approach. *Visual Cognition*, *11*(5), 631–671. https://doi.org/10.1080/13506280344000473

Oksama, L., & Hyönä, J. (2008). Dynamic binding of identity and location information: A serial model of multiple identity tracking. *Cognitive Psychology*, *56*(4), 237–283. https://doi.org/10.1016/j.cogpsych.2007.03.001

Patterson, M. D., Isaacson, D. R., Mendonca, N. L., Neogi, N. A., Goodrich, K. H., Metcalfe, M., Bastedo, B., Metts, C., Hill, B. P., DeCarme, D., et al. (2021). An initial concept for intermediate-state, passenger-carrying urban air mobility operations. *AIAA Scitech 2021 Forum*. https://doi.org/10.2514/6.2021-1626

Peißl, S., Wickens, C. D., & Baruah, R. (2018). Eye-tracking measures in aviation: A selective literature review. *The International Journal of Aerospace Psychology*, *28*(3-4), 98–112. https://doi.org/10.1080/24721840.2018.1514978

Politowicz, M. S., Chancey, E. T., Buck, B. K., Unverricht, J., & Petty, B. J. (2023). MPATH (Measuring Performance for Autonomy Teaming with Humans) ground control station: Design approach and initial usability results. *AIAA SciTech 2023 Forum*. https://doi.org/10.2514/6.2023-2525

Ponsoda, V., Scott, D., & Findlay, J. M. (1995). A probability vector and transition matrix analysis of eye movements during visual search. *Acta Psychologica*, *88*(2), 167–185. https://doi.org/10.1016/0001-6918(95)94012-Y

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision*, *3*(3), 179–197. https://doi.org/10.1163/156856888X00122

Radun, J., Nuutinen, M., Antons, J.-N., & Arndt, S. (2017). Did you notice it?—how can we predict the subjective detection of video quality changes from eye movements? *IEEE Journal of Selected Topics in Signal Processing*, *11*(1), 37–47. https://doi.org/10.1109/JSTSP.2016.2607696

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163. https://doi.org/10.2307/271063

Russo, A., Cardoso Junior, M., & Villani, E. (2025). Eye-tracking analysis to assess the mental load of unmanned aerial system operators: Systematic review and future directions. *The Aeronautical Journal*, *129*(1338), 529–558. https://doi.org/10.1017/aer.2024.122

Schieber, F., & Gilland, J. (2008). Visual entropy metric reveals differences in drivers' eye gaze complexity across variations in age and subsidiary task load. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *52*(23), 1883–1887. https://doi.org/10.1177/154193120805202311

Shiferaw, B., Downey, L., & Crewther, D. (2019a). A review of gaze entropy as a measure of visual scanning efficiency. *Neuroscience & Biobehavioral Reviews*, *96*, 353–366. https://doi.org/10.1016/j.neubiorev.2018.12.007

Shiferaw, B. A., Crewther, D. P., & Downey, L. A. (2019b). Gaze entropy measures detect alcohol-induced driver impairment. *Drug and Alcohol Dependence*, *204*. https://doi.org/10.1016/j.drugalcdep.2019.06.021

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science*, *6*(3), 291–298. https://doi.org/10.1177/1745691611406923

Wickens, C. D., Helleberg, J., Goh, J., Xu, X., & Horrey, W. J. (2001). Pilot task management: Testing an attentional expected value model of visual scanning. *Savoy, IL, UIUC Institute of Aviation Technical Report*.

Wickens, C. D., McCarley, J. S., & Gutzwiller, R. S. (2022). *Applied attention theory*. CRC Press. https://doi.org/10.1201/9781003081579

Yang, M., Wang, J., Gao, Y., & Hu, B. (2024). Aim where you look: Eye-tracking-based UAV control framework for automatic target aiming. *IEEE Internet of Things Journal*, *11*(12), 21250–21260. https://doi.org/10.1109/JIOT.2024.3390115

Zuber, B. L., Stark, L., & Cook, G. (1965). Microsaccades and the velocity-amplitude relationship for saccadic eye movements. *Science*, *150*(3702), 1459–1460. https://doi.org/10.1126/science.150.3702.1459