AFRL-AFOSR-UK-TR-2022-0109

Sensitivity of the spatial distribution of fixations to variations in the type of task demand and its effectiveness as a trigger for adaptive automation

Di Nocera, Francesco
UNIVERSITA' DEGLI STUDI DI ROMA LA SAPIENZA
PIAZZALE ALDO MORO 5
ROMA, ROMA, 00185
ITA

08/26/2022
Final Technical Report

Air Force Research Laboratory
Air Force Office of Scientific Research
European Office of Aerospace Research and Development
Unit 4515 Box 14, APO AE 09421

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
20220826	Final	START DATE 20180315	END DATE 20220314

4. TITLE AND SUBTITLE
Sensitivity of the spatial distribution of fixations to variations in the type of task demand and its effectiveness as a trigger for adaptive automation

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
	FA9550-18-1-0203	
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)
Francesco Di Nocera

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
UNIVERSITA' DEGLI STUDI DI ROMA LA SAPIENZA
PIAZZALE ALDO MORO 5
ROMA, ROMA 00185
ITA

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
EOARD UNIT 4515 APO AE 09421-4515	AFRL/AFOSR IOE	AFRL-AFOSR-UK-TR-2022-0109

12. DISTRIBUTION/AVAILABILITY STATEMENT
A Distribution Unlimited: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT
The proposed research activity investigates the spatial distribution of eye fixations as a real-time measure of mental workload. Recent studies have successfully related the distribution of eye fixations to the mental load. The scope of this research project is to devise a set of experiments for separating the contribution of three types of tasks demands (i.e., temporal, mental, and physical) and to determine which of these (and when) should be considered for using an index of spatial distribution as a trigger in ocular-based adaptive systems.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT U	b. ABSTRACT U	c. THIS PAGE U	SAR	52

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
NANDINI IYER	314-235-6161

Sensitivity of the spatial distribution of fixations to variations in the type of task demand and its effectiveness as a trigger for adaptive automation

Francesco Di Nocera & Piero Maggi

Department of Psychology, Sapienza University of Rome

1. INTRODUCTION

Ocular activity is known to be sensitive to variations in mental workload and many attempts have been made to derive a stable measure of the cognitive resources allocated to a task using information provided by eye-trackers. Recent studies carried out in our laboratory have successfully related the distribution of eye fixations to the mental load and the scope of the research activity described in this report was to separate the contribution of three types of task demands (cognitive, temporal, and physical) and for determining which of these (and when) should be taken into consideration for using an index of spatial clustering as trigger in ocular-based adaptive systems.

More specifically, the aim of this research was threefold: 1) assessing the sensitivity of the proposed measure to different types of task demands with a large sample and a within-subject design; 2) assessing the effectiveness of the proposed measure as a trigger for adaptive automation and 3) extending the spatial analysis of the scanpath using more complex algorithms.

Given the basic nature of the research activity presented here, a simple visuo-motor task was used in laboratory experiments. Although these effects could be extended to operational settings, a testing activity in complex/realistic settings is out of the scope of the present research.

Nevertheless, the research activity described here aimed at investigating the use of the spatial distribution of eye fixations as a real-time measure of mental workload and -therefore- as trigger for adaptive systems. This approach can be easily implemented in all transportation domains (see Di Nocera et al., 2020), not to mention all the operational settings in which an operator is sitting in front of a display (e.g. control rooms). This approach was initially introduced by Di Nocera, Camilli & Terenzi (2007), and recent evidence from secondary sources confirmed that the spatial distribution of eye fixations is sensitive to variations in mental workload (e.g. Chen *et al.,* 2022; Dillard et al., 2014; Fidopiastis et al., 2009; Foy & Chapman, 2018). Earlier studies on the functional significance of this index and its sensitivity to different task demands suggested that fixations appear to be spreaded out when task load depends on the temporal demand, whereas fixation clustering seems to depend on the visuo-spatial demand (Camilli, Terenzi & Di Nocera, 2008). The Nearest Neighbor Index (NNI) used in that research programme showed several competitive advantages over other ocular indicators

usually invoked as measures mental workload: 1) it provides information over the entire scanpath on the examined visual scene instead of depending on pre-defined Areas of Interest (as it happens with “entropy”); 2) it can be computed over relatively small epochs (1-minute), thus making it possible to obtain ongoing information about the functional state of an individual; 3) it is based on published research and can be computed using open access tools, unlike other patented tools such as the Index of Cognitive Activity by Marshall (2007). Moreover, the NNI doesn’t need of post-hoc analyses (as it happens with Event Related Potentials) and it is highly employable in real-world settings, making it a good candidate measure of the mental load for triggering some automated systems adaptively.

Here further investigated the diagnosticity of the index by manipulating the type of demand (mental, temporal, physical) imposed to the individual, compared it to the entropy approach, and attempted to extend it by ...

2. STUDY 1

The objective of this experiment was to test the sensitivity and diagnosticity (see Wierwille & Eggemeier, 1993) of the NNI, that is how the changes in the visual exploration strategy due to different workload levels and different types of demand imposed by the task are captured by the distribution of fixations.

2.1 Methods

Experimental software development. The Tetris game used in this study was coded using JavaScript and GoogleScript. The gaming area consisted of 300 cells deployed on a grid of 15 columns by 20 rows. Each tetromino (piece) was randomly extracted from a pool composed of 7 different tetrominoes types, and it descended at a constant speed. With the aim of creating three experimental conditions, specific variables have been modified to induce a different type of task demand. In Condition 1, the speed of falling pieces has been manipulated to generate time pressure (temporal demand); in Condition 2, the direction of pieces has been reversed to increase mental demand (each piece appears in the lower part of the game area and then rises to the top); in Condition 3, the interaction with pieces was occasionally blocked, therefore forcing the user to press the control keys several times to move the pieces (physical demand).

The manipulations were coherent with the NASA-TLX definition of mental, temporal, and physical demand. Mental demand: "How much mental and perceptual activity was required? Was the task easy or demanding, simple or complex?"; Temporal demand: "How much time pressure did you feel due to the pace at which the tasks or task elements occurred? Was the pace slow or rapid?"; Physical demand: "How much physical activity was required? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?". Therefore, we consider the visuo-spatial demand imposed here as an expression of the mental demand.
In our version of Tetris, the "game-over" consisted of the exhaustion of the playing area given by the excessive accumulation of pieces but did not represent the end of the game. When the event occurs, the program automatically resets the entire area deleting all the accumulated pieces and allowing the user to continue the game until the end of the experiment. The number

of pieces accommodated and the number of completed lines were used as performance measures. The number and shape of the pieces, the size of the playing area, and the difficulty between levels were based on the original version of the Tetris.

Ocular activity recordings. The Gazepoint GP3HD eye-tracking system was used to record ocular activity. This system allows the researcher to collect ocular data without using invasive and/or uncomfortable head-mounted instruments. Gazepoint, the eye tracker manufacturer, claims accuracy within 0.5 to 1.0 degrees and reads data at a rate of 150Hz. The eye tracker was calibrated using the default 9-point calibration test using Gazepoint's included software.

Participants. Thirty university students (19 women and 11 males, M = 25 years old, SD = 3.6) volunteered and participated in the experiment. All participants had normal or corrected-to-normal vision and were naïve as to the aims of the experiment. This research study was completed with the tenets of the Declaration of Helsinki and was approved by the Institutional Review Board of the Department of Psychology, Sapienza University of Rome. Informed consent was obtained from each participant. Participants received a € 20.00 worth bookstore gift card. One subject was excluded from the data analysis due to the low quality of recorded eye movements.

Procedure. Participants were tested in a within-subject design in which the same task was manipulated -in three different sessions- for manipulating the mental, the temporal, and the physical demand. Participants played a custom-coded version of the Tetris game, a commonly known tile-matching puzzle videogame successfully used in a variety of studies (e.g., Trimmel & Huber, 1998). For experimental purposes, the game restarted from a blank screen each time the stack of Tetrominoes reached the top of the gaming area and no new Tetrominoes were able to enter. This condition commonly denotes the end of the game, whereas, in this experiment, it was scored as a loss (performance measure). Participants were instructed to gain as many points as possible (i.e., complete lines and avoid losses).

Training session. Before the experimental session started, each participant performed a training session whose scope was to familiarize the participants with the experimental setting. To this aim, each participant played the Tetris game starting from a low difficulty level and moving on to Baseline, Temporal demand (TD), Mental demand (MD), and Physical demand (PD) conditions. The training had a 5-minute duration and did not include the evaluation of the participants' performance level in this phase. The scheme of the training session is reported below:

- One minute of gameplay at Level 1 (drop speed: 1250 ms per block), with the aim of verifying the correct understanding of the game rules and allowing the participant to familiarize with the use of directional keys.
- Baseline condition: one minute, configured at level 6 (drop speed: 208 ms per block). It was used to acquire the baseline for the experimental session.
- TD condition: one minute, set at level 8 (drop speed: 156 ms per block).

- MD condition: one minute, during which the entire playing area was rotated by 180°, and each Tetromino appeared on the bottom side and went up, accumulating on the top of the gaming area.
- PD condition: one minute, in which the participant needed to press the directional keys repeatedly to move the piece quickly in the chosen direction (instead of keeping the key pressed).

The difficulty level of the conditions, as determined by the drop speed of the pieces, was defined on the basis of previous studies (Camilli, Terenzi & Di Nocera, 2008; Camilli, Terenzi & Di Nocera, 2007).

Experimental session. After the calibration of the eye-tracker, participants were instructed to play the game earning as many points as possible (i.e., complete lines and avoid losses). Each condition lasted 10 minutes, and the order of presentation was randomized across participants. After completing each condition (Baseline vs. TD vs. MD vs. PD), participants were requested to fill in the NASA-TLX (Hart & Staveland, 1988).

2.2. Data analysis and results

Performance data. A performance index (PI) was computed based on the number of lines completed in relation to the maximum number of lines that could be completed. The maximum value is obtained by the total number of Tetrominoes that the participant managed in each condition (For example, with 60 pieces, it was possible to complete a total of 16 lines if managed in an optimal way). The index goes from 0 to 1, where 1 means that the player has obtained the maximum achievable score. The performance index was used as the dependent variable in a repeated-measures ANOVA design, using Condition (Baseline vs. TD vs. MD vs. PD) as a repeated factor. Results showed a main effect of the condition [$F_{3, 84} = 16.88$, $p < .001$] (Figure 2). The faster (TD) and Reversal conditions (MD) were associated to the worse performance with respect to the baseline (Table 1). Overall, performance is low in all conditions. This could indicate either a limited ability of the subjects in the Tetris game or a wrong selection of difficulty levels (described in the previous paragraph). However, the baseline is easier than the experimental conditions (excluding the PD condition), which allows us to read the results obtained from the subjective and ocular measures in line with the starting hypotheses.

	TD	MD	PD
Baseline	.001	.011	.285
TD		.01	.001
MD			.001

Table 1. Post-hoc analysis carried out through the Duncan test. Pairwise comparison among PI scores and conditions.

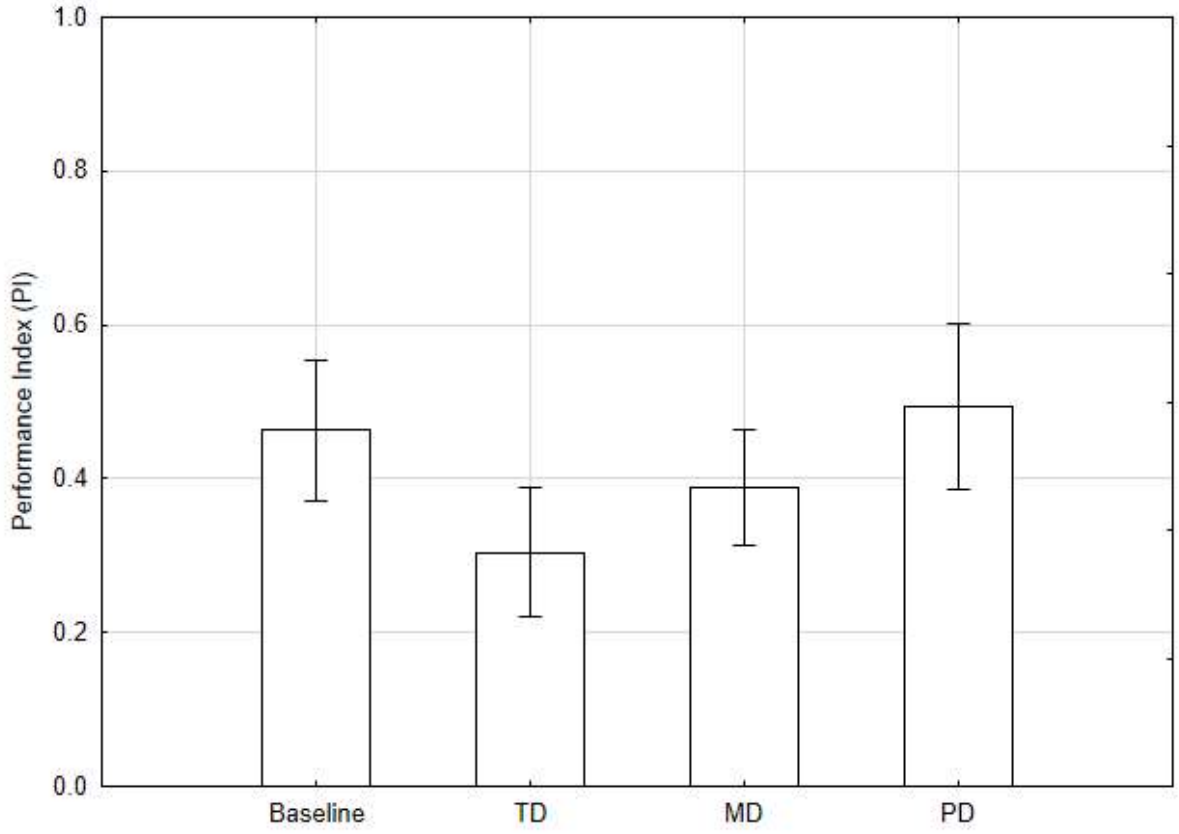

Figure 2. The performance index shows the overall strategy used by participants. It is the ratio of the number of completed lines to the number of pieces that appeared during the game. Values close to 1 mean an optimal game with a high number of lines completed.

Subjective measure. NASA-TLX weighted ratings were used as dependent variables in a repeated-measures ANOVA design using Condition as repeated factor. Results showed a main effect of Condition [$F_{3, 84} = 11.11$, $p < .001$] (Figure 3, Table 2), consistent with those obtained for the performance index. Although analyses on the single items are questionable from a statistical standpoint, it is worth noting that TD, MD, and PD conditions showed higher values for temporal, mental, and physical demand scales, respectively (Figure 4, Table 3-5). These results show that the manipulations made with the Tetris have indeed taxed specific aspects or resources.

	TD	MD	PD
Baseline	.001	.001	.51
TD		.153	.001
MD			.01

Table 2. Post-hoc analysis carried out through the Duncan test. Pairwise comparison among NASA-TLX scores and conditions.

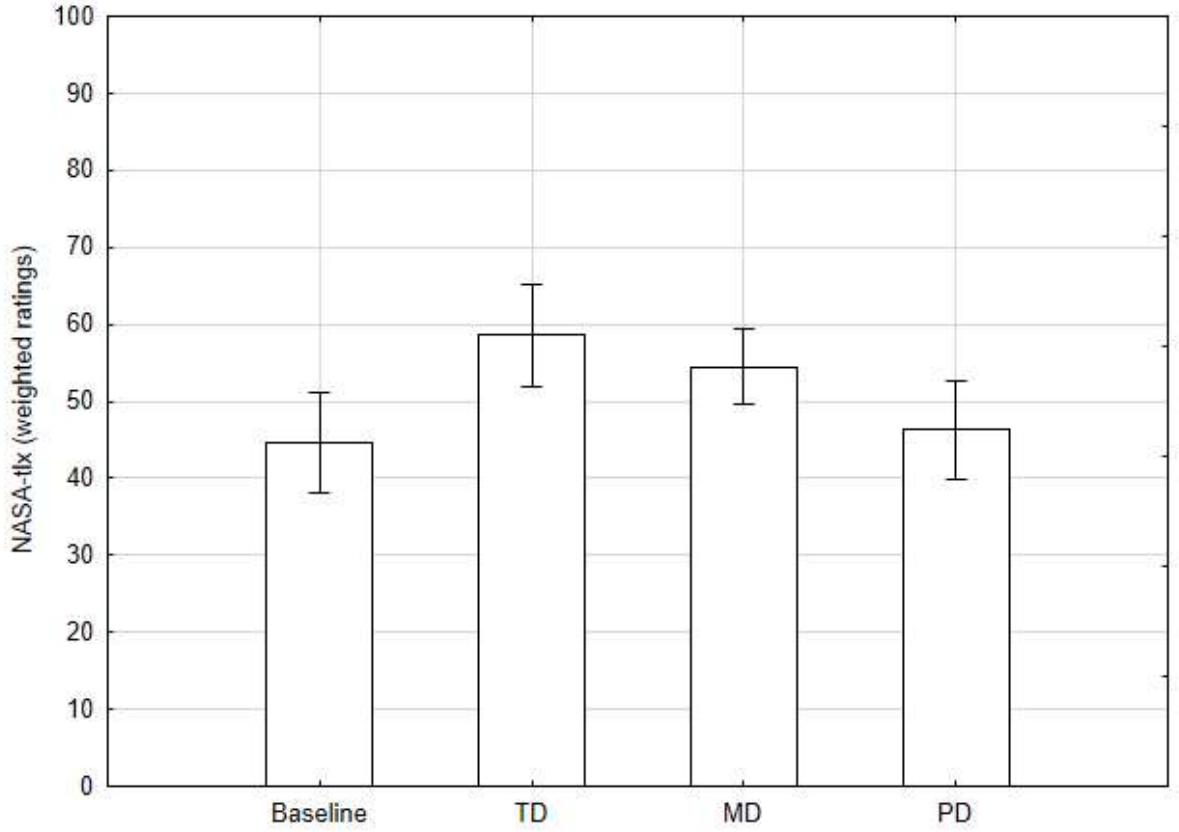

Figure 3. NASA-TLX values (weighted scores) separately for the conditions. Error bars denote .95 confidence intervals.

	TD	MD	PD
Baseline	.001	.829	.917
TD		.001	.001
MD			.765

Table 3. Post-hoc analysis carried out through the Duncan test. Pairwise comparison among Temporal demand scale scores of NASA-TLX and conditions.

	TD	MD	PD
Baseline	.05	.001	.612
TD		.01	.016
MD			.001

Table 4. Post-hoc analysis carried out through the Duncan test. Pairwise comparison among Mental demand scale scores of NASA-TLX and conditions.

	TD	MD	PD
Baseline	.882	.66	.001
TD		.583	.001
MD			.001

Table 5. Post-hoc analysis carried out through the Duncan test. Pairwise comparison among Physical demand scale scores of NASA-TLX and conditions.

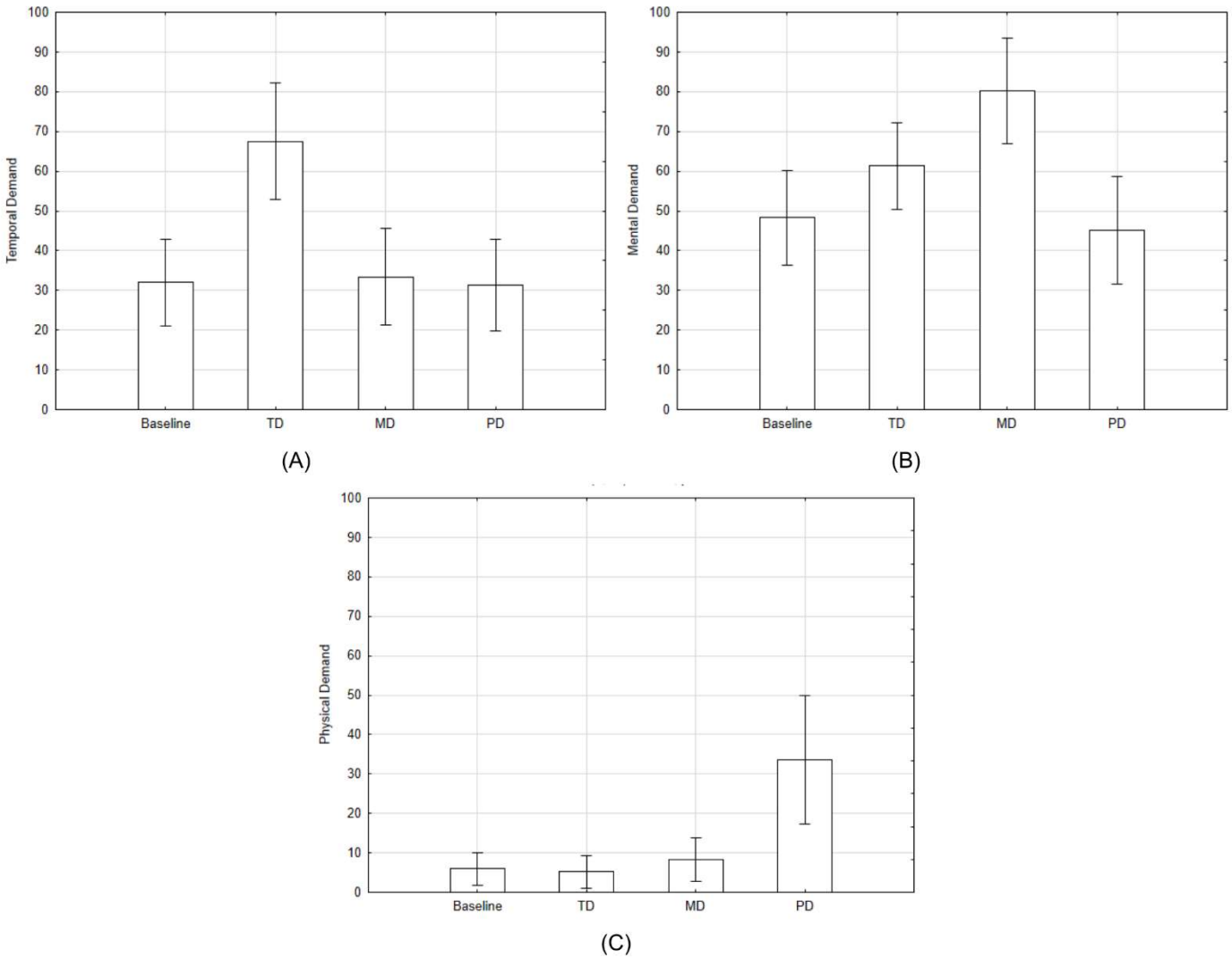

Figure 4. NASA-TLX subscales values (Temporal (A)[$F_{3, 84} = 14.36$, $p < .001$], Mental (B)[$F_{3,84} = 12.685$, $p < .001$] and Physical (C) [$F_{3,84}=13.27$, $p < .001$] demand) separately for the conditions. Error bars denote .95 confidence intervals.

Number and duration of fixations. The number and duration of fixations were computed on epochs of 1 minute for each participant and then averaged. Averaged number and duration of fixations were used as dependent variables in a repeated-measures ANOVA design using Condition as the repeated factor. No significant differences between conditions were found (Figure 5-A and 5-B; Tables 6-7) [Fixations number: $F_{3, 84} = .365$, $p > .05$] [Fixations duration: $F_{3, 84} = .798$, $p > .05$].

	TD	MD	PD
Baseline	.36	.78	.61
TD		.49	.64
MD			.79

Table 6. Post-hoc analysis carried out through the Duncan test. Pairwise comparison among fixations number and conditions.

	TD	MD	PD
Baseline	.48	.46	.71
TD		.18	.69
MD			.30

Table 7. Post-hoc analysis carried out through the Duncan test. Pairwise comparison among fixations duration and conditions.

Amplitude of saccades. The amplitude of saccades was computed on epochs of 1 minute for each participant and then averaged. The averaged amplitude of saccades was used as dependent variable in a repeated-measures ANOVA design using conditions as a repeated factor (Figure 5-C). Results showed a main effect of condition [$F_{3, 84} = 12.84$, $p < .001$].

	TD	MD	PD
Baseline	.60	.001	.19
TD		.001	.08
MD			.001

Table 8: Post-hoc analysis carried out through the Duncan test. Pairwise comparison among amplitude of saccades and conditions.

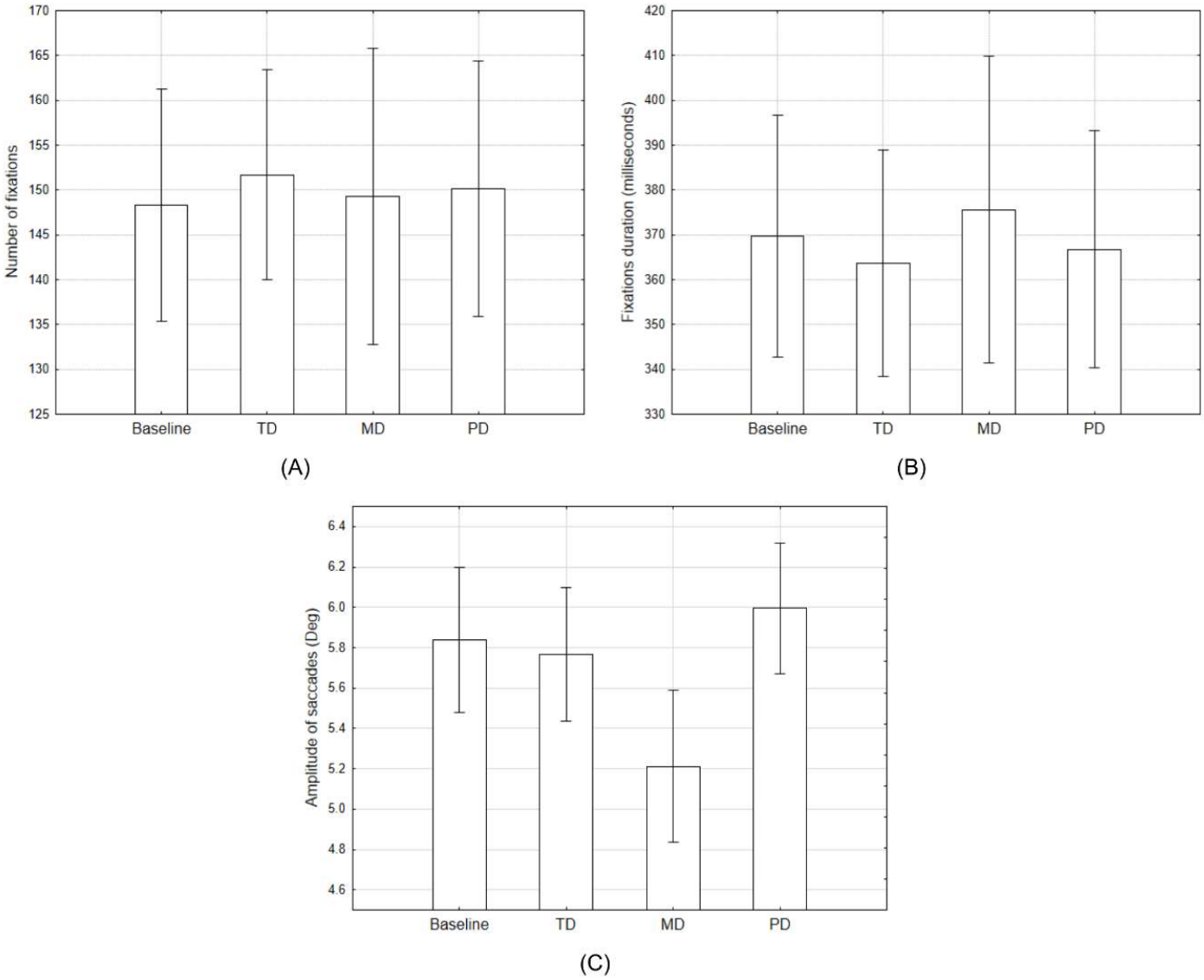

Figure 6: Averaged number (A) and duration (B) of fixations, and amplitude of saccades (C), for the conditions compared with the baseline. Error bars denote .95 confidence intervals.

Nearest Neighbor Index (NNI). The NNI was computed on epochs of 1 minute (see Di Nocera, Ranvaud & Pasquali, 2015 for detailed explanation) for each participant and then averaged. Averaged NNI values were used as the dependent variable in a repeated measures ANOVA using conditions as the repeated factor. Results showed a main effect of Condition [$F_{3, 84} = 12.31$, $p < .001$]. TD condition showed higher NNI values (i.e., a more dispersed distribution of fixations) than the baseline (Figure 7-A), while in the MD condition, NNI values were the lowest (Figure 6-B) (Table 9).

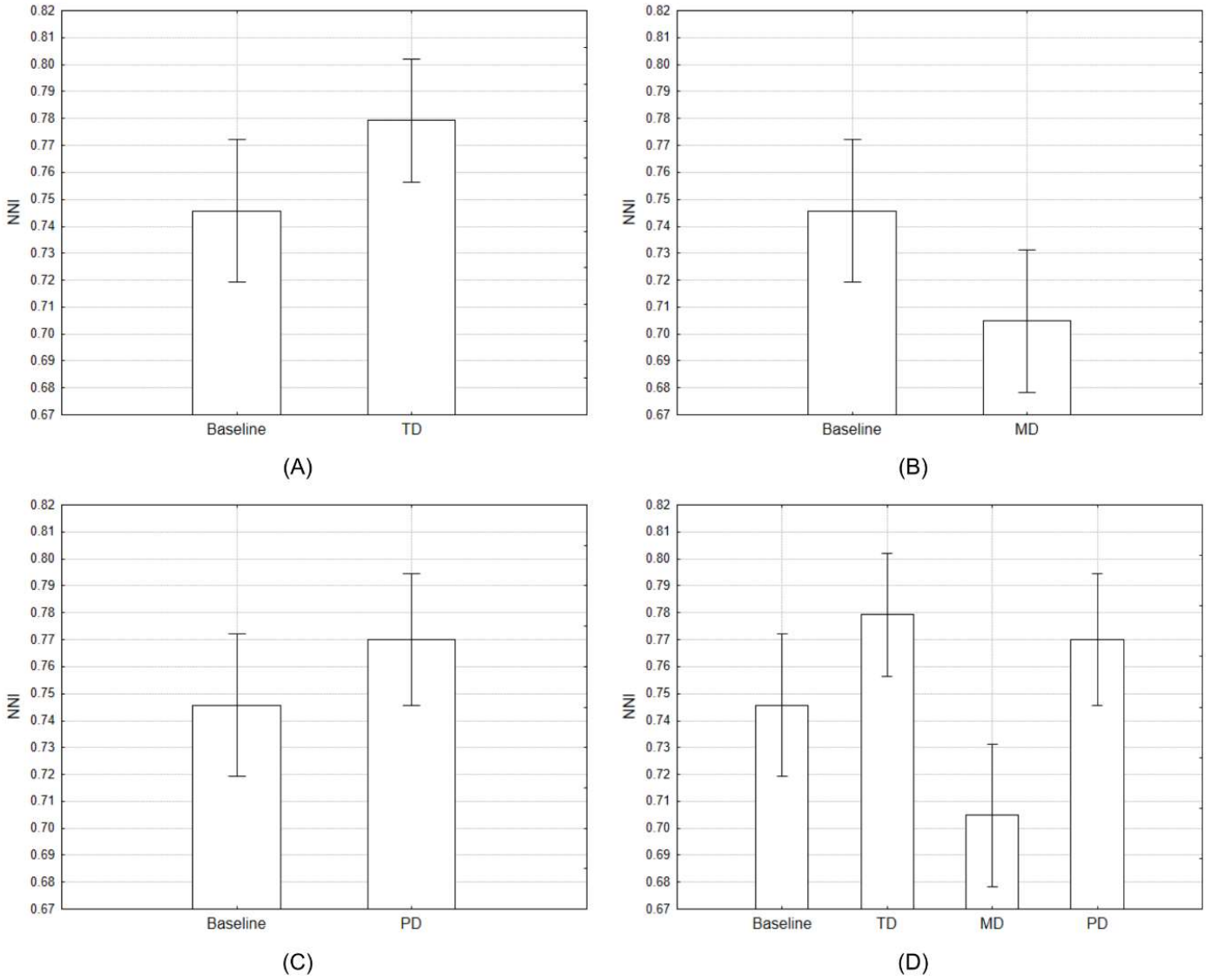

Figure 6. Average NNI for the conditions compared with the baseline separately. Error bars denote .95 confidence intervals. Baseline Vs TD (A); Baseline Vs MD (B); Baseline Vs PD (C); all condition (D).

	TD	MD	PD
Baseline	.018	.01	.072
TD		.001	.49
MD			.001

Table 9. Post-hoc analysis carried out through the Duncan test. Pairwise comparison among NNI and conditions.

2.3. Discussion

This first study aimed at investigating how the visual exploration strategy changes both along with the task load and with the type of task demand. The results showed an increase in the NASA-TLX values of the single subscales (mental demand, physical demand, and temporal demand) matching the respective manipulation. Overall, we observed a greater workload in the MD and TD conditions compared to the control and PD conditions. The latter has shown higher values in the corresponding NASA-TLX scale, but the manipulation of the physical demand did not affect the overall self-reported workload. Finally, and more important to our aims, the analysis of the fixations pattern showed high clustering when the task-load increment was obtained by changing the mental (visuo-spatial) demand and low clustering when it was obtained by changing the temporal demand.

3. STUDY 2

In this study we perform a direct comparison of two indices of mental workload based on the analysis of ocular data: entropy and distribution of fixations.

Entropy can be defined as a measure of the disorder found in any physical system and this concept was then applied by Tole et al. (1983) to eye movements. When the individual looks at all the quadrants in the scene and crosses all the potential combinations of stimuli with a stable frequency, the entropy will increase. Instead, the entropy value will be lower when the individual focuses attention on a narrower range of possible areas of interest. That happens because the frequency of transitions from one area to another decreases. A regular and systematic visual exploration strategy is shown in a condition of low entropy, which corresponds to a more orderly passage to other areas. The principal benefit of this analysis is the possibility to "summarize" the visual strategy using a single value. The index is indicated in bits/second.

The Nearest Neighbor Index, instead, provides data on the distribution of points in space. The average distance between the fixations collected during the execution of a task and the average distance between the fixations expected in a random distribution are taken into account in the application of the NNI to eye movements. The result is represented by a single value where one indicates that the empirical and the random distribution are not different; values above one indicate dispersion, while values below one show clustering. The index can be computed for small epochs if sufficient fixations are available (about 50 as a rule of thumb) and then analyzed as a time series, therefore offering information on the temporal variations of distribution of fixation points. A methodological study (Camilli et al., 2007) supports the validity of this algorithm as a measure of mental workload, highlighting the consistency of the index with subjective and psychophysiological measures.

The entropy-based analysis of the scanpath and the spatial distribution of fixations points are reported to be good indices of mental workload. However, they have never been directly compared.

3.1. Methods

Stimuli. To induce high visual-spatial demand and to assess how that affects visual search, a single pair of complex black and white drawings was used (figure 7 and 8). Drawings were rich in details so that the numerous elements would engage participants in a long visual exploration session. The size of each picture was 9.8 x 5.5 inches, featured thirty-five subtle differences but were otherwise identical. The two images were aligned horizontally and in full-screen mode on a 27" display.

Ocular activity recordings. Prior to recording, participants performed a nine-point calibration, and then their eye movements were recorded through the Pupil Labs system with binocular 120 Hz Eye Tracking Camera (Pupil Labs GmbH, Germany) claims accuracy of 0.6 degrees.

Participants. The experiment involved fourteen university students (9 women and 6 males, mean age = 24 years, S.D. = 2.6) who participated on a voluntary basis. All participants had normal or corrected-to-normal vision and were naïve as to the aims of the experiment. This study was compliant with the principles of the Declaration of Helsinki and was the protocol approved by the Institutional Review Board of the Department of Psychology, Sapienza University of Rome. Informed consent was obtained from each participant. Participants received a € 20.00 worth bookstore gift card.

Procedure. The experiment was conducted in a dark room, and participants were seated at approximately 2 ft. from a computer screen. During the task, they had to find as many differences as they could between the two images in a 24-minute session. They were requested to click with the mouse on each difference they identified. The differences found were automatically highlighted with a circle throughout the session. Participants were also asked to provide a subjective evaluation of mental workload on a 2-minute schedule (Instantaneous Self-Assessment (ISA): Tattersall & Foord, 1996).

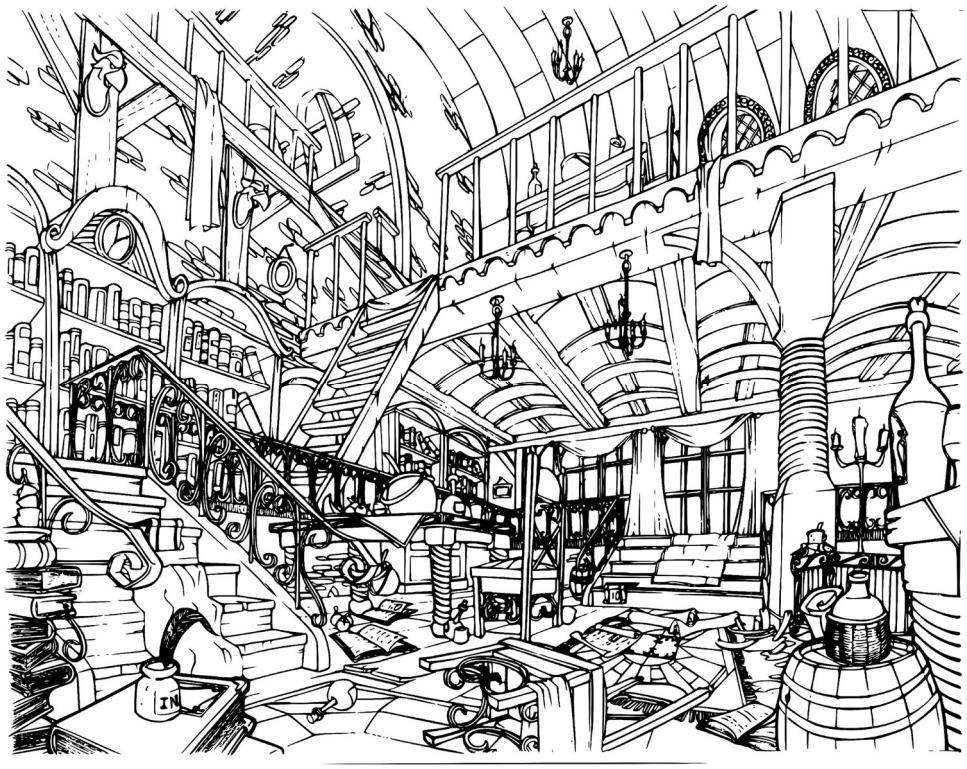

Figure 7. Left panel. Artwork by Benoit Tranchet (reproduced with permission).

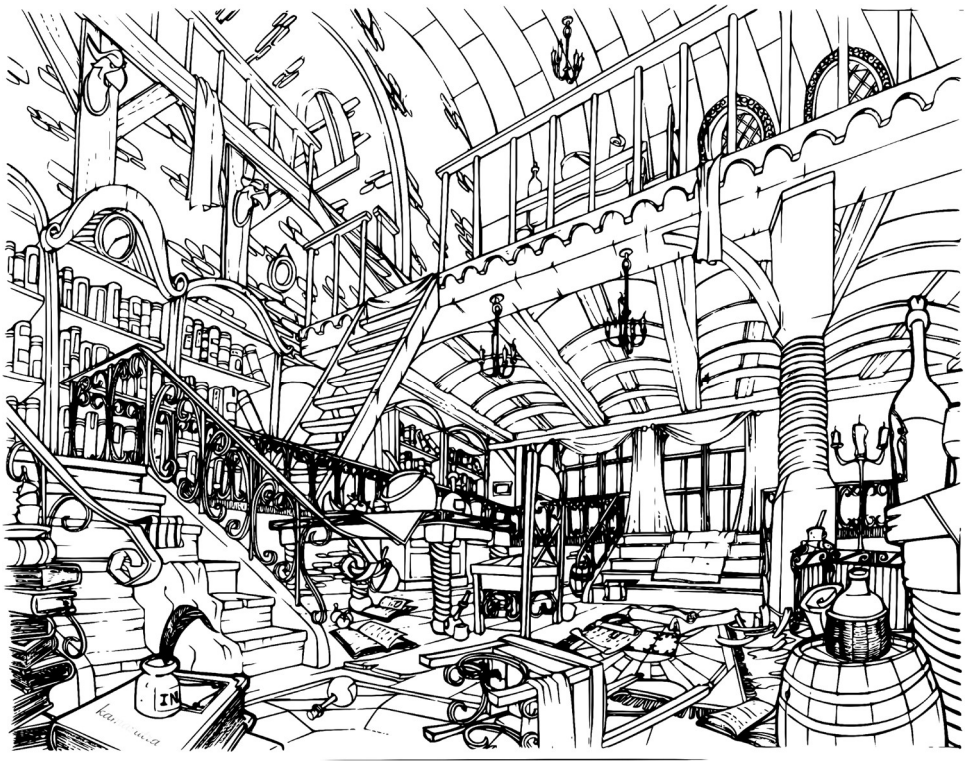

Figure 8. Right panel. A modified version of the original artwork with 35 differences.

3.2. Data analysis and results

Performance and self-report measures. The whole activity was split into 12 periods of two minutes each to match performance and subjective evaluations. The number of differences identified by each subject in each epoch was used as a performance indicator. The number of differences identified and the ISA scores were used as dependent variables in two repeated measures ANOVA designs using Epoch as a repeated factor. A main effect of Epoch was found both on the number of differences [$F_{11, 143} = 16.52$, $p < .001$] (figure 9) and the ISA scores [$F_{11, 143} = 15.50$, $p < .001$] (figure 11). Plots reveal Duncan's post-hoc testing revealed an asymptotic pattern for both the performance measure and the workload estimates starting from the twelfth minute (Tables 10-11).

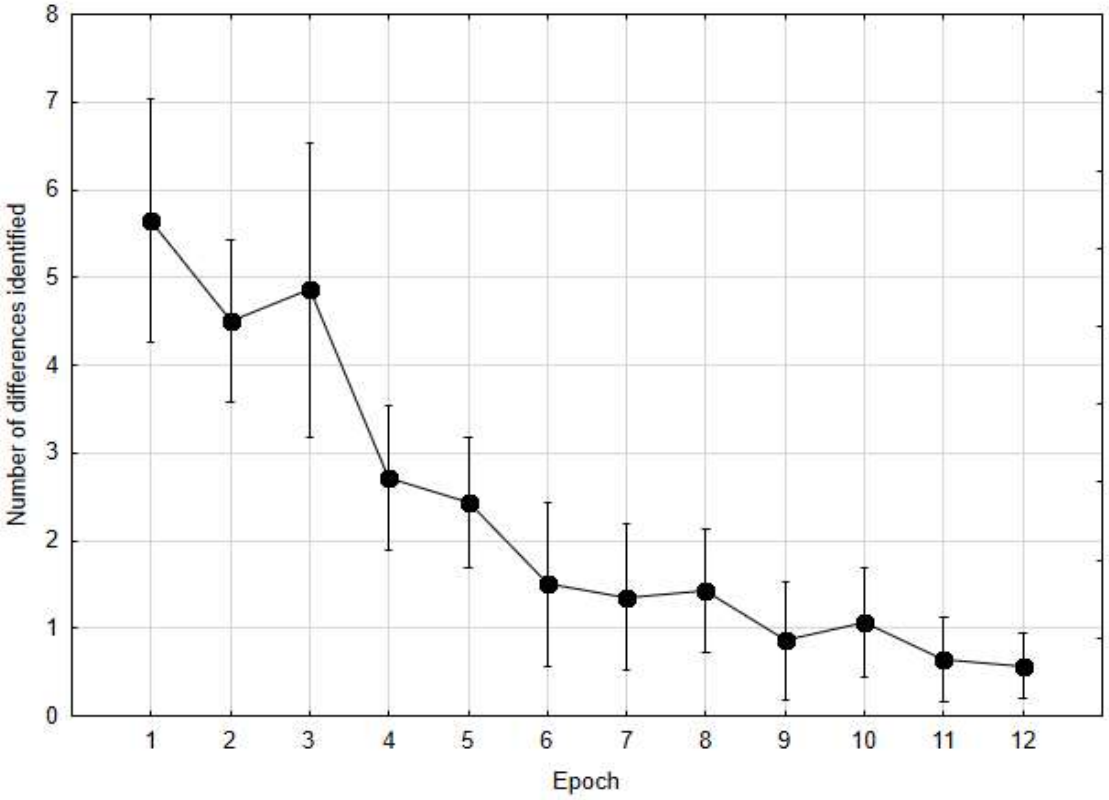

Figure 9. Task performance (number of differences found) along time. Error bars denote .95 confidence intervals.

Epochs	2	3	4	5	6	7	8	9	10	11	12
1	.079	.201	.001	.001	.001	.001	.001	.001	.001	.001	.001
2		.561	.004	.001	.001	.001	.001	.001	.001	.001	.001
3			.001	.001	.001	.001	.001	.001	.001	.001	.001
4				.642	.061	.048	.055	.007	.017	.002	.002
5					.131	.113	.125	.023	.048	.010	.008
6						.829	.908	.362	.533	.231	.201
7							.908	.448	.642	.296	.263
8								.405	.588	.263	.231
9									.728	.728	.665
10										.516	.467
11											.908

Table 10. Post-hoc analysis carried out through the Duncan test. Pairwise comparison among Number of differences identified and Epochs.

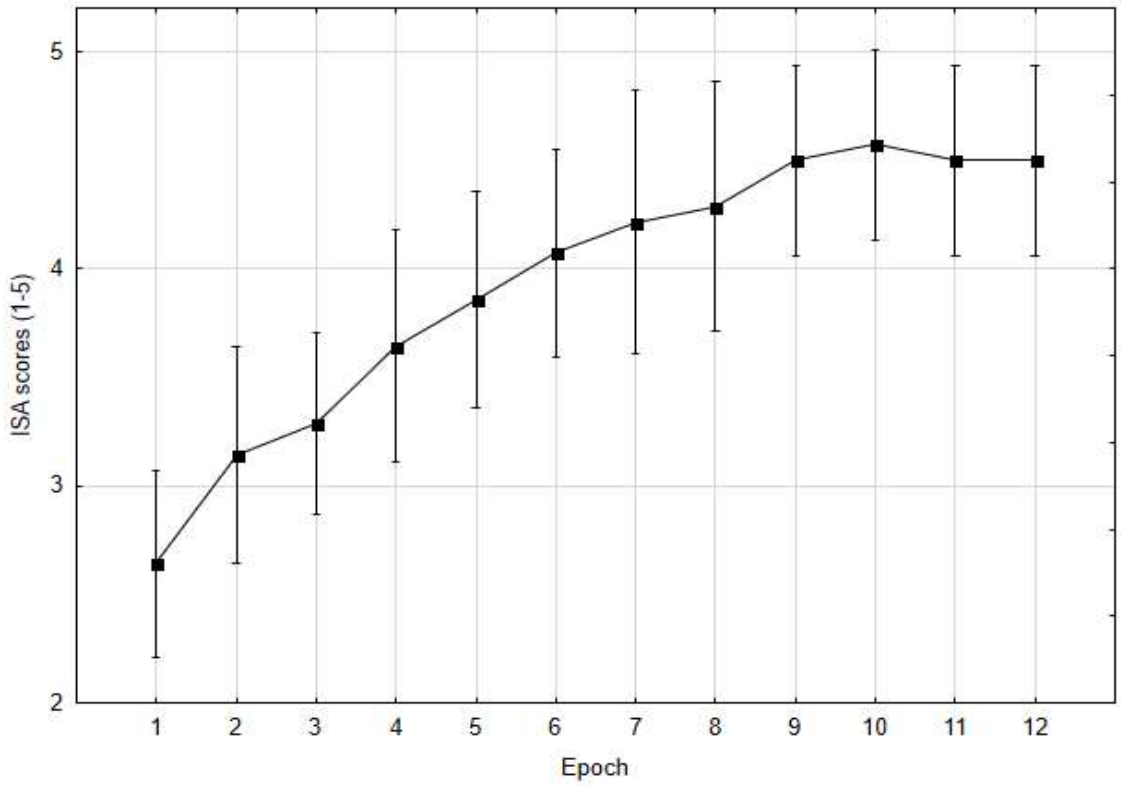

Figure 10. Subjective workload (ratings from 1 to 5) along time. Error bars denote .95 confidence intervals.

Epochs	2	3	4	5	6	7	8	9	10	11	12
1	.028	.006	.001	.001	.001	.001	.001	.001	.001	.001	.001
2		.529	.036	.003	.001	.001	.001	.001	.001	.001	.001
3			.116	.016	.001	.001	.001	.001	.001	.001	.001
4				.345	.074	.019	.009	.001	.001	.001	.001
5					.345	.138	.085	.012	.005	.011	.009
6						.529	.377	.100	.056	.093	.085
7							.753	.270	.174	.257	.238
8								.397	.270	.377	.345
9									.753	1.000	1.000
10										.770	.779
11											1.000

Table 11. Post-hoc analysis carried out through the Duncan test. Pairwise comparison among ISA scores and Epochs.

Nearest neighbor Index (NNI). For each participant, the NNI was calculated taking into account 1-minute epochs. Average NNI values were used as dependent variables in a repeated-measures ANOVA design using Epoch as a repeated factor. A main effect of the Epoch was found [$F_{11, 143} = 4.41$, $p < .001$] (Figure 11). Duncan's post-hoc testing showed that the visual strategy applied in the first two minutes significantly differs from all other periods (Table 12).

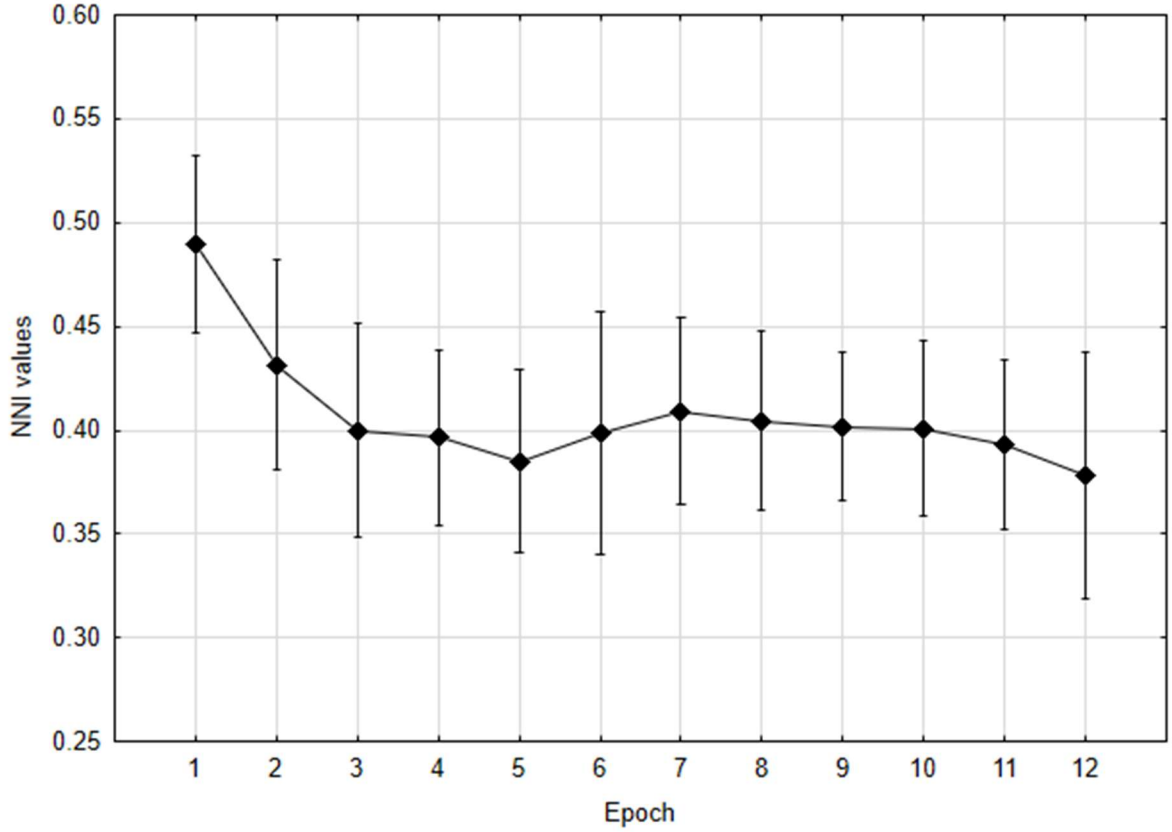

Figure 11. NNI values along time. Error bars denote .95 confidence intervals.

Epochs	2	3	4	5	6	7	8	9	10	11	12
1	.003	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001
2		.163	.131	.046	.154	.253	.197	.172	.165	.099	.021
3			.869	.513	.953	.683	.829	.923	.969	.751	.347
4				.591	.907	.589	.723	.808	.847	.857	.408
5					.536	.315	.411	.474	.501	.692	.728
6						.653	.794	.884	.927	.782	.364
7							.819	.732	.700	.495	.197
8								.890	.850	.618	.269
9									.948	.696	.317
10										.732	.338
11											.487

Table 12. Post-hoc analysis carried out through the Duncan test. Pairwise comparison among NNI scores and Epochs.

Entropy rate. The whole visual area has been divided into two Areas of Interest (AOI), namely the two images displayed. For each minute, the maximum number and duration of fixations made on each AOI were assessed. For these AOIs, the entropy rate has been adopted as a measure of scan randomness (Tole et al., 1983). The entropy rate (H-rate) is expressed in units of bit / s (i.e., the information given by each observation, assessed in bits over seconds). A random pattern is represented by a high H-rate. In this study, all the scanpaths performed by the participants were used to compute the entropy rate. The entropy rates (H_rate) of the sequences of one length for the two images used were computed as a measure of the randomness of the scan. Average H_rate values were used as dependent variables in a repeated-measures ANOVA design using the epoch as a repeated factor. A main effect of time [$F_{11.143} = 3.69$, $p < .001$] was found (Figure 12). Duncan's post-hoc testing showed a steady pattern in the first two minutes of visual exploration, consistently with that obtained with the NNI (Table 13).

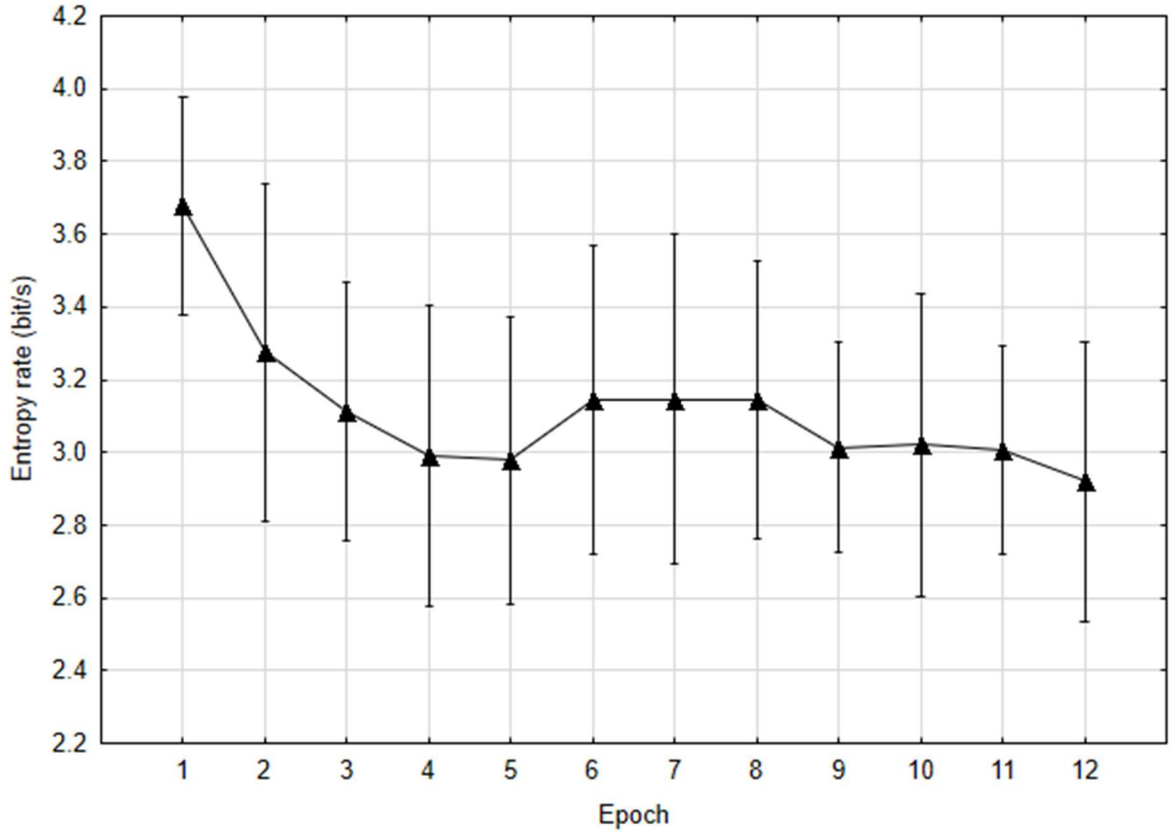

Figure 12. Entropy rate values along time. Error bars denote .95 confidence intervals.

Epochs	2	3	4	5	6	7	8	9	10	11	12
1	.007	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001
2		.337	.110	.097	.432	.412	.381	.132	.137	.129	.046
3			.481	.444	.825	.837	.843	.534	.537	.530	.280
4				.929	.379	.387	.393	.894	.863	.914	.660
5					.346	.353	.358	.836	.807	.854	.702
6						1.000	.999	.426	.434	.422	.209
7							.999	.440	.453	.432	.214
8								.449	.466	.439	.218
9									.960	.972	.593
10										.937	.572
11											.604

Table 13. Post-hoc analysis carried out through the Duncan test. Pairwise comparison among H-rate scores and Epochs.

3.3. Discussion

This second study aimed at comparing two scanpath analysis methods that have been previously reported to be sensitive to changes in the task-load: Entropy rate and Nearest Neighbor Index. Results showed an overall increase of difficulty after the first few minutes of the task. The entropy rate confirms the presence of a less random and more stereotyped pattern starting from the second minute of recording. A similar trend was found for the NNI. The average NNI values in the first two minutes of activity were significantly higher than in the following epochs, therefore showing a change towards fixations grouping as the task-load increased. This study was designed to evaluate the potential of these two measures under the effect of increasing visual-spatial demand. The results showed the same trend, therefore confirming that the two indices are sensitive to changes in the visuo-spatial demand. However, unlike the entropy rate, the NNI is also suitable for estimating changes due to the temporal demand (see Study 1). This is an aspect that could not be accommodated by the entropy rate, which is based on the transitions between AOIs, and on the visuo-spatial performance.

4. STUDY 3

The third study explored the possibility of using the Nearest neighbor Index as a trigger within an adaptive automation system through two steps: i) identifying the right modality of automation, verifying if it is helpful for the individual; ii) observing if NNI values return to the baseline after the implementation of the automation support.

An automation system was embedded in the Tetris version described in the first study. An "autopilot", able to take total control of the system, was designed as the best solution to avoid game-over in critical situations. Therefore, a function has been added to detect alignment errors and calculate the best possible combination. The automatic positioning is done by simulating in the background all combinations between the piece played and those previously (accumulated in the lower part of the area). The system selects the optimal combination, calculated considering the maximum size of the piece surface that touches the bottom of the play (an example is shown in figure 13).

Figure 13. An example of different combinations, computed in background, for the autopilot. The second image represents an incorrect positioning, while in the third image, we see the best combination compared to the one shown in the first image.

4.1. Methods

Experimental design. The design involved the use of two independent variables with two levels, leading to four experimental conditions. The "Difficulty" variable (i.e., Easy, Hard) and the "Automation" variable (i.e., Present, Absent). All subjects performed the 4 conditions in random order (Table 14). For each condition, ocular, performance, and subjective measures were acquired: NNI, lines completed, and NASA-TLX, respectively. Automation consists of the activation of the autopilot that takes control of the game until the player turns it off. In order to keep participants engaged, during the automation mode, they were requested to report by pressing the bar key each time the piece turned white (blinked) for 200 milliseconds. The interval between blinks varies between 3 and 6 seconds. The study aimed at determining whether the NNI can be used as a trigger in an adaptive automation system and whether the autopilot is an optimal automation level for the purpose. Therefore, we have tried to verify this by the following assumptions: i) Subjects will perceive the hard condition as more complex than the easy one. This simply confirms that the two conditions have a different mental workload and ii) The easy condition will not be different from conditions with manual automation.

		Automation	
		Present	Absent
Difficulty	Easy	Condition 1 (EAut)	Condition 2 (EMan)
	Hard	Condition 3 (EAut)	Condition 4 (HMan)

Table 14. Two independent variables at two levels, including four experimental conditions.

Tools and software. The Gazepoint GP3HD eye-tracking system was used to record ocular activity. This system allows the researcher to collect ocular data without using invasive and/or uncomfortable head-mounted instruments. Gazepoint, the eye tracker manufacturer, claims accuracy within 0.5 to 1.0 degrees and reads data at a rate of 150Hz. The eye tracker was calibrated using the default 9-point calibration test with Gazepoint's included software.

Participants. Eighteen university students (10 women and 8 males, mean age = 27.3 years, St. dev. = 3.5) volunteered in the experiment. All participants had normal or corrected-to-normal vision, and they were naïve as to the aims of the experiment, its expected outcomes, and its methodology. This research complied with the tenets of the Declaration of Helsinki and was approved by the Institutional Review Board of the Department of Psychology, Sapienza University of Rome, Italy. Informed consent was obtained from each participant. Participants received a € 20.00 worth bookstore gift card.

Procedure. Four conditions have been created (Table 15):

1. EMan: Easy level of difficulty, without automation
2. HMan: Hard level of difficulty, without automation
3. EAut: Easy level of difficulty, with automation
4. HAut: Hard level of difficulty, with automation

After the eye-tracker's calibration, the subjects were explained that their task was to play the game, earning as many points as possible (i.e., complete lines and avoid losses). Each condition lasted 10 minutes, and the order of presentation was randomized across participants. In conditions where manual automation was present, subjects received instruction to activate it by pressing the "CTRL key" any time they perceived a too high difficulty level. Then, the autopilot took control of the game until the subject deactivated it using the same input (i.e., CTRL key). To keep the subject engaged during the autopilot execution, a secondary detection task was asked to be performed: in this phase, the piece, controlled by the computer, turned white for 200ms at intervals between 3 and 6s. The task consisted of pressing the spacebar as soon as possible every time that happened.
After completing each condition, participants were requested to fill in the NASA-TLX (Hart & Staveland, 1988).

4.2. Data analysis and results

Performance data. To analyze the performance between different conditions, we computed a Performance Index (PI). The PI was based on the number of lines completed in relation to the maximum number of lines that can be performed. The maximum value is obtained by the total number of pieces that the subject managed in each condition (For example, with 60 pieces, it is possible to complete a total of 16 lines if managed in an optimal way). The index goes from 0 to 1, where 1 means that the player has obtained the maximum achievable score. The performance values were used as dependent variables in a two-factor repeated-measures ANOVA design, using difficulty level (easy or hard) and present/absent automation as factors. The interaction effect was not significant [$F_{1, 17} = .389$, $p > .05$]. However, the results showed a main effect of the Difficulty [$F_{1,17} = 25.37$, $p < .001$] and Automation factors [$F_{1,17} = 15.86$, $p < .001$]. The conditions with the presence of automation were associated with high performance with respect to the absence of automation (Figure 14). Moreover, according to the assumptions, better performance is observed under easy conditions than under the difficult ones (Figure 15).

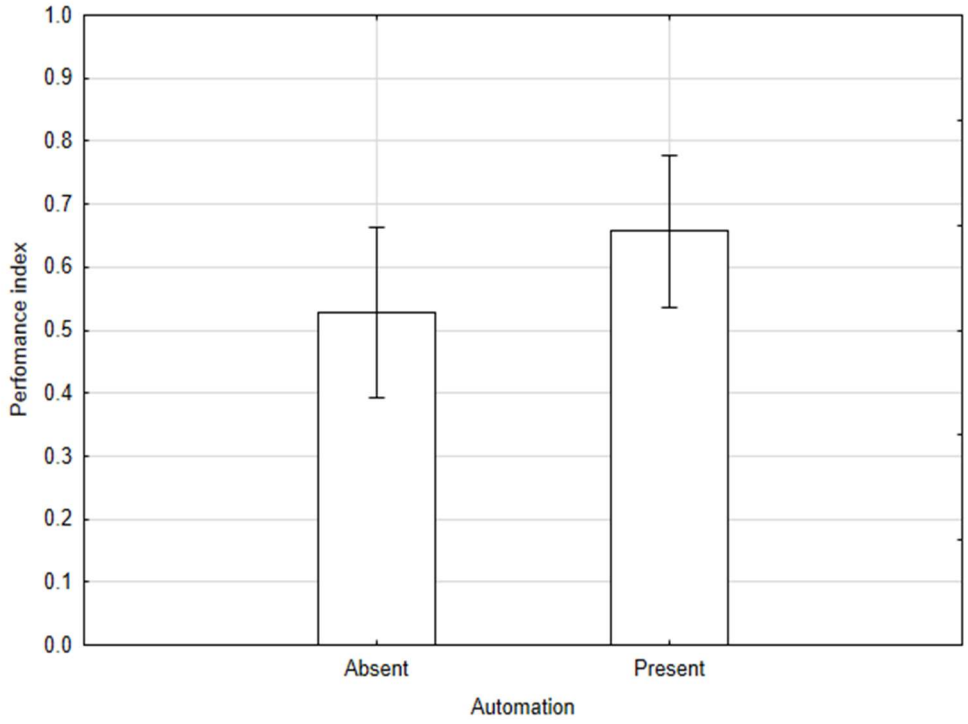

Figure 14. Performance index, the main effect of automation factor. Error bars denote .95 confidence intervals.

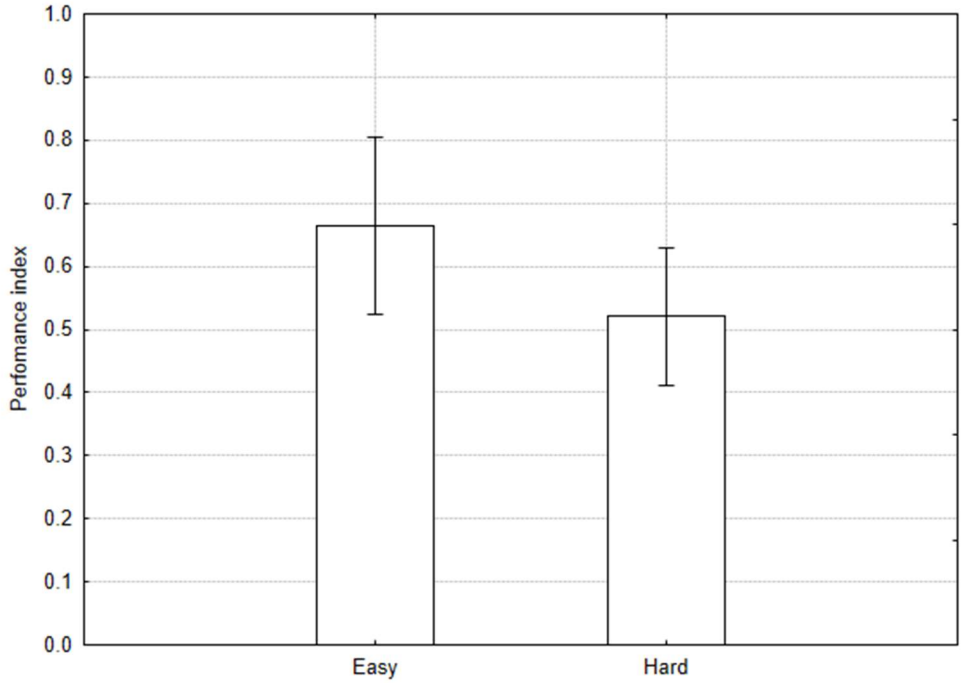

Figure 15. Performance index, the main effect of the difficulty factor. Error bars denote .95 confidence intervals.

Subjective measure. NASA-TLX weighted ratings were used as dependent variables in a two-factor repeated-measures ANOVA design, using difficulty level (easy or hard) and present/absent automation as factors. The results did not show an effect of interaction [$F_{1, 17}$ =.68, p > .05]. We observed a significant main effect of the Difficulty factor [$F_{1,17}$ = 22.73, p < .001] (Figure 16), consistent with the performance results. However, there are no differences between conditions with and without Automation [$F_{1, 17}$ = .152, p > .05]. The latter suggests that automation has not changed the perception of difficulty provided by the subjects, although the objective performance values are consistent with the assumptions made.

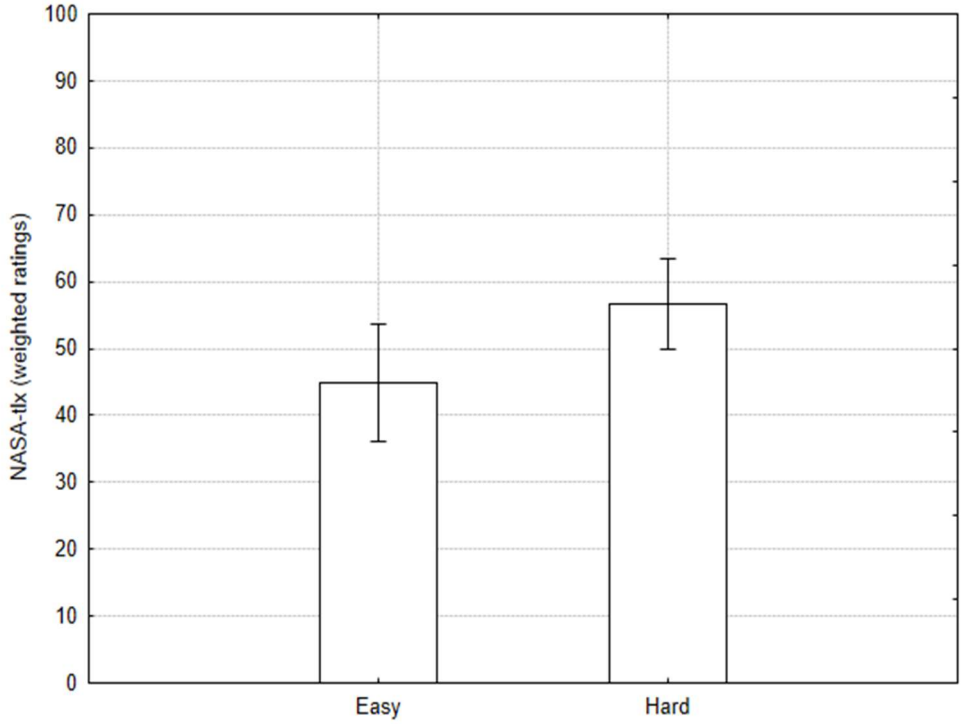

Figure 16. NASA TLX score, the main effect of the difficulty factor. Error bars denote .95 confidence intervals.

Nearest Neighbor Index. The NNI was computed on epochs of 1 minute for each participant. Averaged NNI values were used as dependent variables in a two-factor repeated-measures ANOVA design, using difficulty level (easy or hard) and present/absent automation as factors. The results showed a significant interaction effect between difficulty and automation factors [$F_{1,17} = 14.78$, $p < .01$] (Figure 17). When Automation was introduced, NNI values were not different between difficulty levels. Moreover, these ratings are very similar to those related to the EMan condition. The hard condition, without automation, results in the highest NNI values (Table 15).

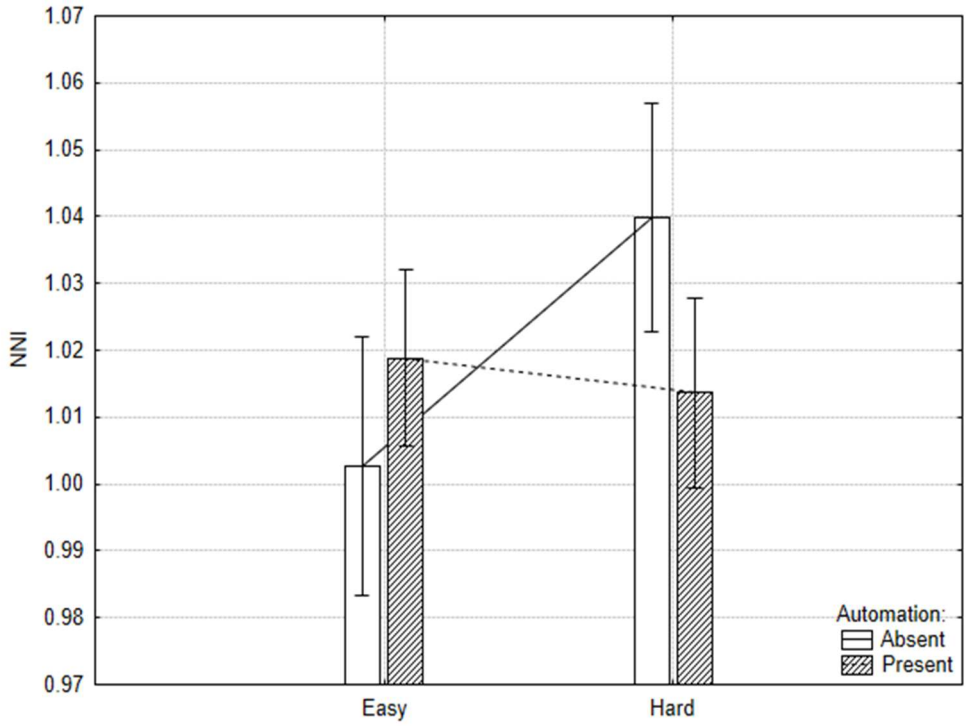

Figure 17. NNI values, the interaction effect between difficulty and automation factors. Error bars denote .95 confidence intervals.

	HMan	EAut	HAut
EMan	.001	.064	.177
HMan		.015	.004
EAut			.515

Table 15. Post-hoc analysis carried out through the Duncan test. Pairwise comparison among NNI scores and Conditions.

The NNI points were categorized using a threshold value obtained from the EMan condition. This threshold was calculated per subject, and it considered the average plus one standard deviation of the NNI scores in the "Easy-Without Automation" condition. Subsequently, for

all conditions, the total number of values that exceeded this limit was calculated. This analysis is based on the idea that the NNI points set up a range of "Normality" in the optimal condition. Therefore, in the non-optimal conditions, the scores can be expected outside this range, indicating a change in the subject's visual exploration strategy during a critical situation. NNI scores over the threshold, which is defined in a version of the task with an optimal difficulty level, could be used as powerful triggers in an automation system. The total number of NNI values over the threshold were used as dependent variables in a two-factor repeated-measures ANOVA design, using difficulty level (easy or hard) and present/absent automation as factors. The results showed a significant interaction effect between difficulty and automation factors [$F_{1,17} = 9.07$, $p < .01$] (Figure 18). In difficult conditions, without automation (HN), the results show the highest number of trigger values than EN and HMA conditions. However, HN was not different from the EMA condition (Table 16).

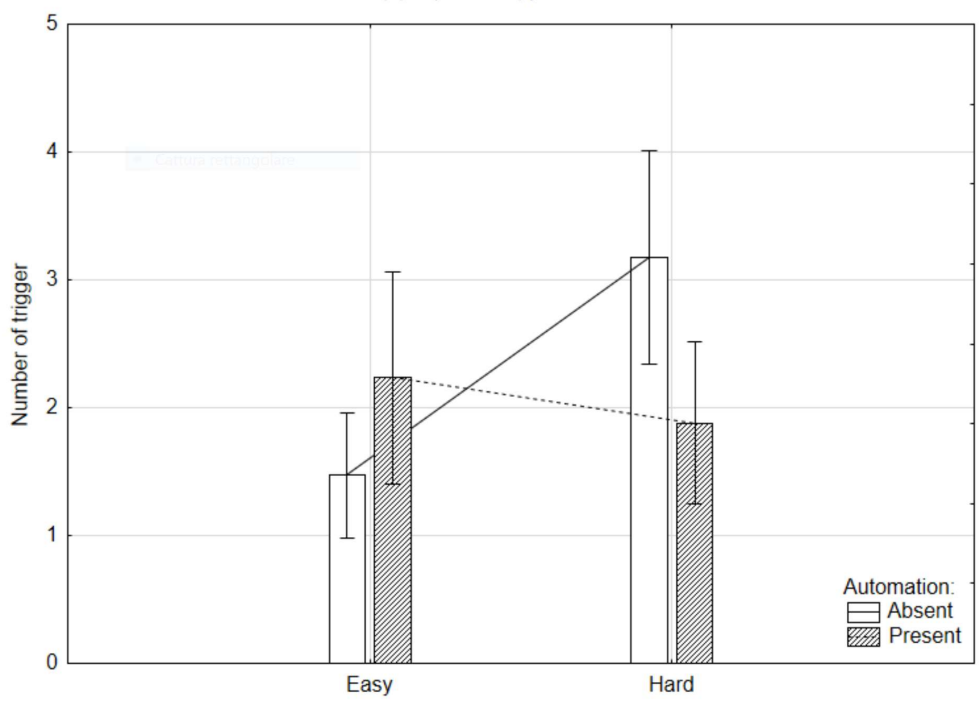

Figure 18. The total number of NNI values over the threshold (Number of triggers), the interaction effect between difficulty and automation factors. Error bars denote .95 confidence intervals.

	HMan	EAut	HAut
EMan	.001	.151	.406
HMan		.068	.02
EAut			.475

Table 16. Post-hoc analysis carried out through the Duncan test. Pairwise comparison among Number of triggers and Conditions.

4.3. Discussion

This study aimed to test the effectiveness of the autopilot as an automation system. Therefore, performance, subjective and ocular measures were compared among different conditions. All measurements confirm a higher level of difficulty in the HMan condition than in the EMan. The analyses on NNI and performance values showed an interesting trend. Regardless of the difficulty level, automation conditions are not harder than the easiest without automation, which was set to be an optimal game condition. However, the subjective workload assessment is not consistent with this result. Here, the hard conditions (with and without automation) do not show significant differences.

On the one hand, automation seems to be an effective aid in terms of performance. On the other hand, it does not seem to affect the perception of the overall difficulty. This can be explained by the frequent switching between manual and autopilot gaming. Not surprisingly, frequent switching between automation levels could increase mental workload.

5. STUDY 4

The study aims to validate NNI as a trigger in an adaptive automation system. The experiment did not provide a real "adaptivity", rather the automation was activated/deactivated according to a schedule defined a priori by the investigator, and that varies for each subject. Therefore, a series of “3-min units” was defined, allowing us to observe what happens, in terms of ocular strategies, before and after the present/absence of automation. The effectiveness of the automation itself was verified, which consists of the autopilot described in study 3, through subjective measurements (i.e., NASA-TLX), performance (i.e., the number of lines performed), and ocular metrics (NNI).

5.1. Methods

Participants. Thirty university students (i.e., 17 women and 13 males, mean age = 26.3 years, St. dev. = 3.2) volunteered in the experiment. All participants had normal or corrected-to-normal vision, and they were naïve as to the aims of the experiment, its expected outcomes, and its methodology. This research complied with the tenets of the Declaration of Helsinki and was approved by the Institutional Review Board of the Department of Psychology, Sapienza University of Rome, Italy. Informed consent was obtained from each participant. Participants received a € 20.00 worth bookstore gift card.

Apparatus. The Gazepoint GP3HD eye-tracking system was used to record ocular activity. This system allows the researcher to collect ocular data without using invasive and/or uncomfortable head-mounted instruments. Gazepoint, the eye tracker manufacturer, claims accuracy within 0.5 to 1.0 degrees and reads data at a rate of 150Hz. The eye tracker was calibrated using the default 9-point calibration test with Gazepoint's included software.

Procedure. The experiment included two game sessions of 10 and 31 minutes, respectively. In the first 10-minute session, a baseline of eye movements was calculated, thus defining the threshold given by the average NNI values ±1 standard deviation. Subsequently, it was used to identify the periods of high complexity (when the NNI values exceeded this threshold). The 10 minutes were divided into two 5-minute units, one with automation and the other without random order. As in previous studies, the entire session was set to an easy level of play, level 6: with a drop speed of 208 ms per block. In the second 31-minute phase, automation was activated/deactivated based on a schedule set by the experimenter and customized for each subject. The schedule was created using a partial randomization method: each 31-min session contained a total of fifteen "3-minute trials" (defined as "series"), where the third minute corresponds to the first of the next series. Only half of the series included the autopilot's use, provided only in the second minute of each one (figure 19). This session was set to a hard level of play (level 8: with a drop speed of 156 ms per block) to increase the mental workload. At the end of each phase (the first and last 5 minutes of the training session and the 31-minute session), NASA-TLX was administered.

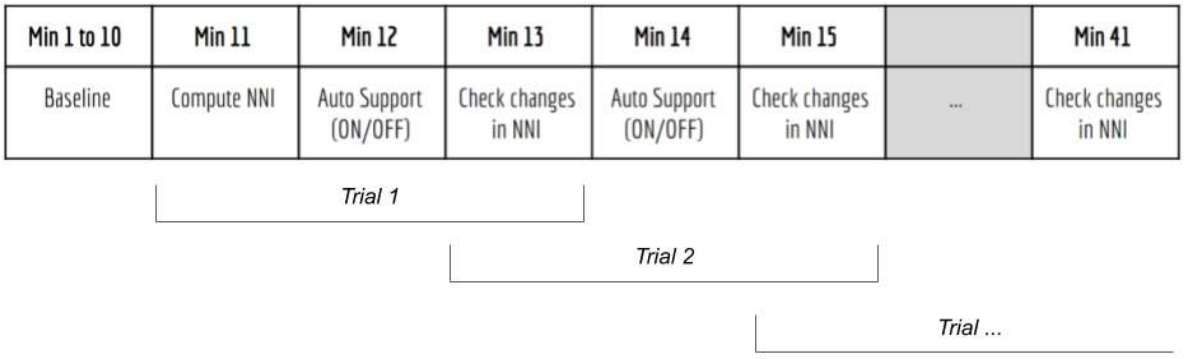

Figure 19. Graphic representation of the setting.

5.2. Data analysis and results

Performance data. The Performance Index (PI) was calculated as in previous studies. The performance values were used as dependent variables in a two-factor repeated-measures ANOVA design, using sessions (training and "31 minutes" sessions) and present/absent automation as factors. The interaction effect was significant [$F_{1, 29} = 40.85$, $p < .001$]. The results showed that subjects achieve lower performance when automation is absent. In addition, there is a difference between the training session and the experimental session (i.e., 31 minutes: "31-min"), again when automation is absent (Table 17). Subjects play better during the training phase, which is set to be easier, according to the assumptions above (better performance is observed under easy conditions (Figure 20).

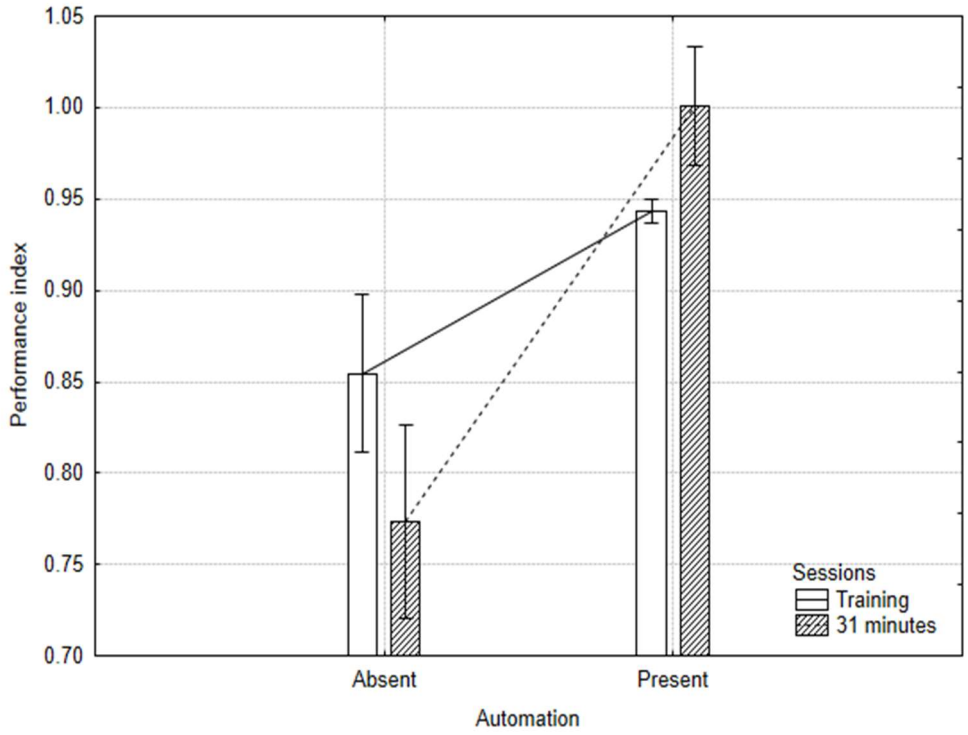

Figure 20. Performance index, the interaction effect between sessions, and automation factors. Error bars denote .95 confidence intervals

	Training (Auto)	31-min (NoAuto)	31-min (Auto)
Training (NoAuto)	.001	.001	.001
Training (Auto)		.001	.001
31-min			.001

Table 19. Post-hoc analysis carried out through the Duncan test. Pairwise comparison among PI score and Conditions.

Subjective measures. NASA-TLX weighted ratings were used as dependent variables in repeated measures ANOVA design using three conditions as the repeated factor: First and last 5-minutes of the training session and the 31-min session. Results showed a main effect of the condition [$F_{2, 58} = 104.78$, $p < .001$] (Figure 21; Table 20), consistent with the performance index and previous studies.

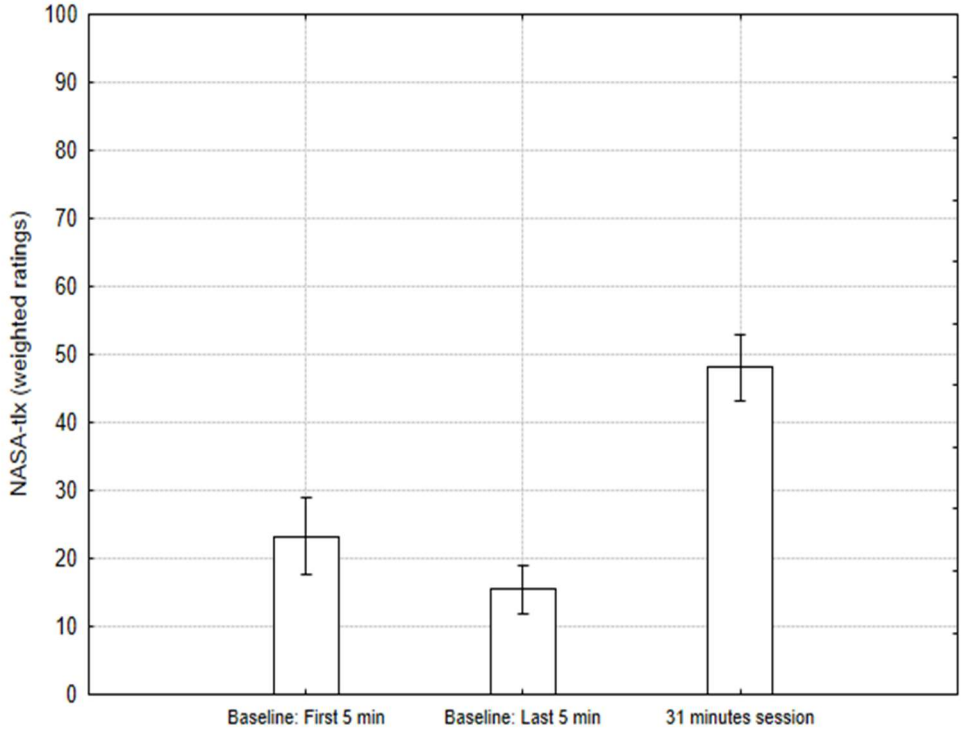

Figure 21. NASA-TLX values (weighted scores) separately for the conditions. "Baseline: First 5 min" refers to the first 5 minutes of the training session, while "Baseline: Last 5 min" refers to the last minutes of the same. Error bars denote .95 confidence intervals.

	Training (Auto)	31-min
Training (NoAuto)	.001	.001
Training (Auto)		.001

Table 19: Post-hoc analysis carried out through the Duncan test. Pairwise comparison among NASA-TLX score and Conditions.

To keep the subject engaged during the autopilot execution, a secondary detection task was performed: in this phase, the piece, controlled by the computer, turned white for 200ms at

intervals between 3 and 6 seconds. The task consisted of pressing the spacebar as soon as possible every time this happened. The values of average reaction times were used as dependent variables in a repeated-measures ANOVA design, using conditions with automation (last 5-minutes of training and "31 minutes" sessions) as repeated factors. Results showed a main effect of the condition [$F_{1, 29} = 93.49$, $p < .001$], the reaction times increased in the 31-min session with respect to the training session with automation (last 5-minutes) (Figure 22).

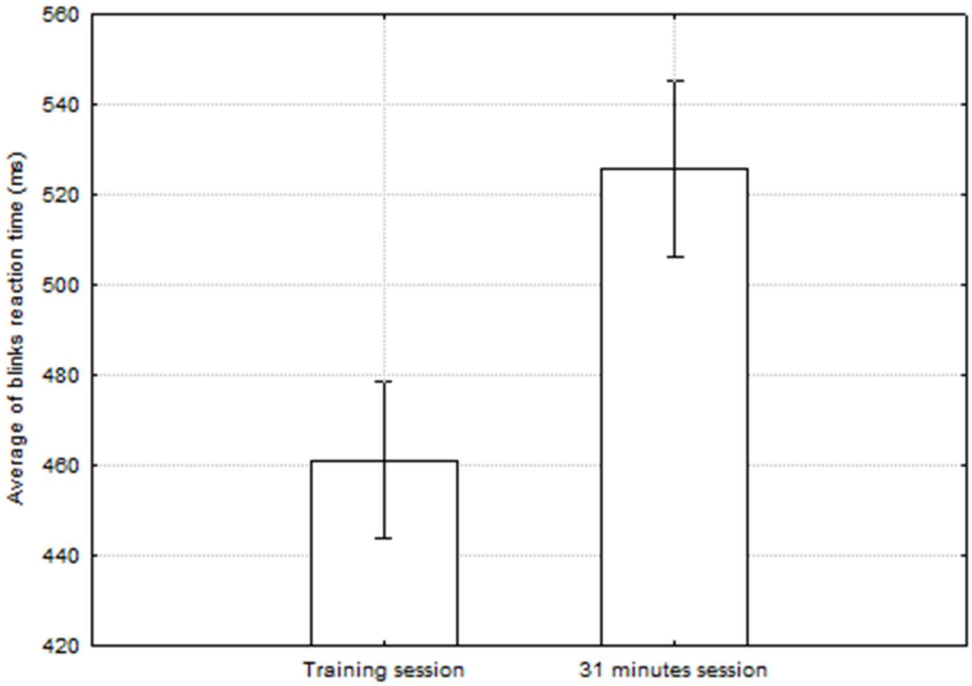

Figure 22. Comparison of average reaction times between the training phase and the 31-minute experimental session. Error bars denote .95 confidence intervals.

Ocular metrics. The presence or absence of automation support was classified into four categories:

- Adaptively **(NP)**: in the first minute of the series, the NNI was out of range, and, in the next minute, the automation was present.
- Invalidly **(nNP)**: in the first minute of the series, the NNI was in the range and, in the next minute, the automation was present.
- Not provided when needed **(NA)**: In the first minute of the series, the NNI was out of range, and, in the next minute, the automation was absent.
- Not provided when not needed **(nNA)**: in the first minute of the series, the NNI was in the range and, in the next minute, the automation was absent.

The proportion of within-range values is expected to be significantly higher in the valid conditions (NP and nNA) rather than in the invalid conditions (NA and nNP). The NNI values obtained in the 31-minute session were used to catalog the single series in NP, nNP, NA, nNA categories. Those four conditions were used as factors in a repeated measure ANOVA design. The results did not show significant differences [$F_{3,81} = 1.35$, $p > .05$] (Figure 23). Furthermore, the comparison between valid and invalid conditions, as two different groups, did not show

significant differences [$F_{1, 29}$ =3.56, p = .069] (Figure 24). This result is not in line with the initial assumptions but shows a compatible trend, and it is possible that more trials may decrease data variability and provide stronger results.

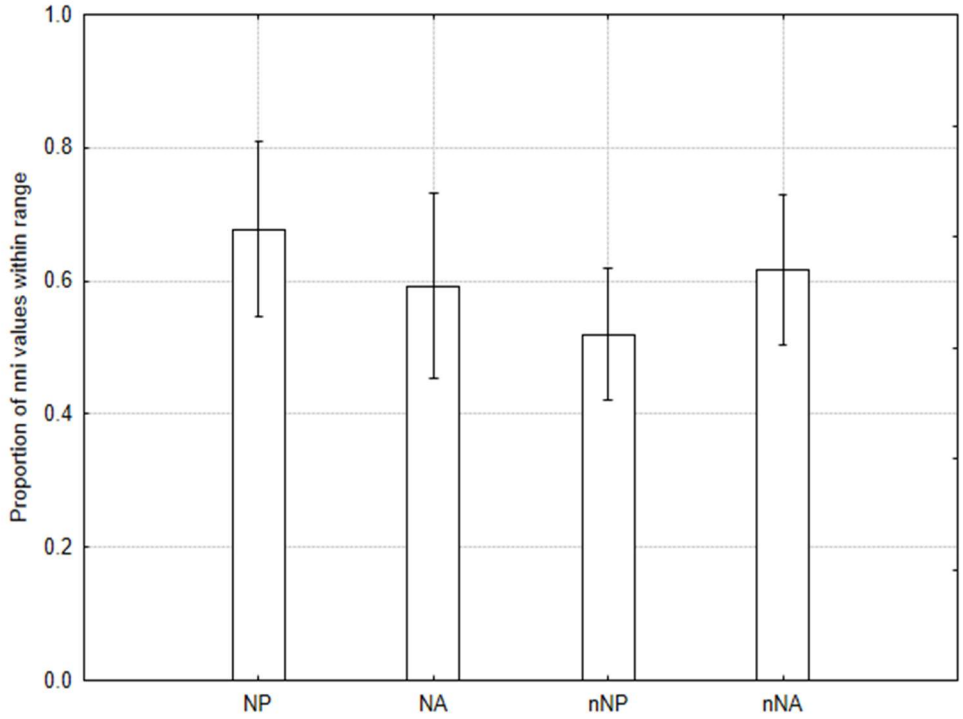

Figure 23. Comparison of NP, NA, nNP, nNA conditions. Error bars denote .95 confidence intervals.

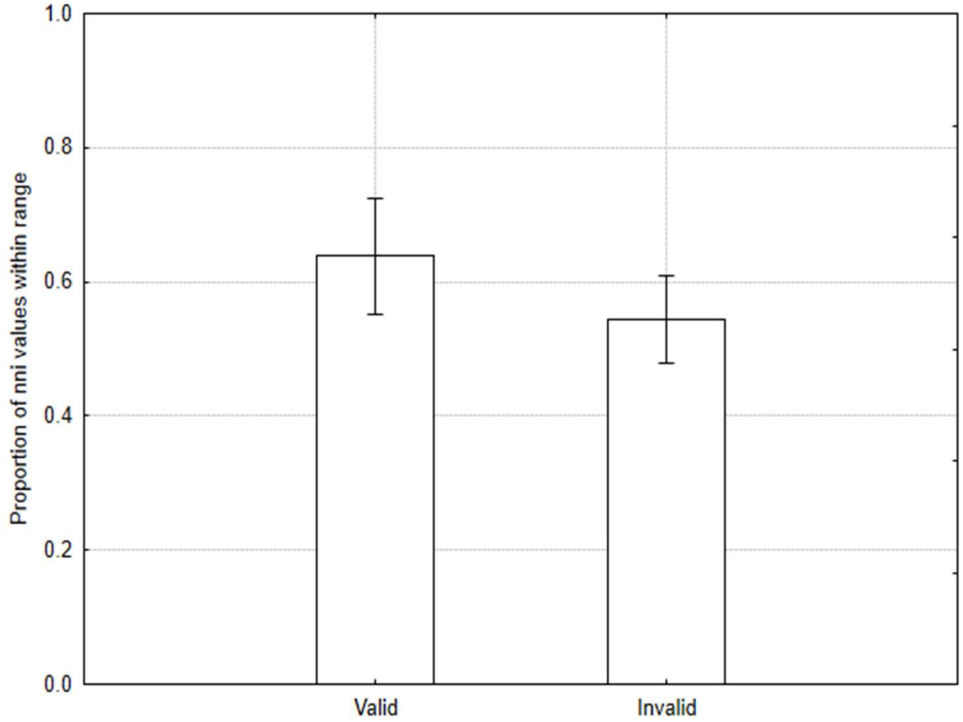

Figure 24. Comparison of conditions Valid (NP, nNA) and Invalid (NA, nNP). Error bars denote .95 confidence intervals.

5.3. Discussion

In this study, the autopilot was applied to compare its effect on eye behavior among different trials. In particular, the 31 minutes were divided into 15 trials, each composed of 3 minutes of detection: in the first minute, the autopilot was absent. In the second minute, this could be present or absent. The presence of automation in the second minute of each trial was random, and therefore, it was presented after a game session in which the subject had shown a high mental workload, operating as it would be adaptive automation. In other cases, it was presented after a game session in which the subject had not shown a high mental workload (NNI within the baseline range), not operating as adaptive automation. The results showed high variability, probably caused by the low number of occurrences for each category (NP, nNP, NA, nNA) and the high dynamism of the task, where a single error can lead to game-over. However, it should be noted that this experiment was based on the calculation of a 10-minute baseline that allowed us to calculate threshold values. Future studies should consider a more durable baseline to provide more accurate values for each subject.

6. ALTERNATIVE METHODS FOR ANALYZING THE SCANPATH

This last section describes some attempts to extend the analysis of the spatial distribution of fixations using other algorithms. Data from Study 1 were re-analyzed to this aim. Algorithms like the Nearest Neighbor Index rarely consider the temporal dimension and sequentiality of points in a trajectory. Linking spatial variation to eye movements over time has been done by determining the distribution of fixations separately in each temporal period (on the minute) (Di Nocera, Ranvaud & Pasquali, 2015). Also, the sequence of eye movements can be analyzed by comparing graphical representations of scanpaths. Di Nocera & Bolia (2007) analyzed pilots' scanpaths using stochastic PERT networks to gather detailed information on the processes underlying the ocular activity. One of the goals of this researchactivity is to link human performance to spatio-temporal patterns in eye-movement data. The number of statistics that could be used (and have never been used) in the spatiotemporal analysis of the scanpath is very large (Stark & Young, 1981; Smith, 1998). For example, in 1975, Pinder and Witherick proposed an adaptation of the NNI algorithm for linear one-dimensional situations. Unlike Clark & Evans (1954) original study, the authors do not consider the area occupied by the points in space but the line that connects them.

The first goal was to obtain a more stable NNI measure over time that can filter out the peaks due to fixations far from the interaction area, often caused by sampling errors or elements outside the task’s area that catch the subject's attention. The previous analyses have always considered the average of more values, showing how this varies with the mental workload. In an application context, each minute's value should be the result of the NNI net of the "noise" mentioned above.

Next, the NNI was analyzed in the frequency domain. This second goal focused on examining the NNI from a new perspective, trying to detect "outlier" frequencies in comparison with frequencies generated by an optimal condition. To realize it, the data from the first study were reused as the basis for the new analyses.

K-Nearest neighbor algorithm. The k-nearest neighbors (k-NN) is an algorithm used to classify objects based on the characteristics of items near the targeted one (Figure 25). An item is classified based on the majority vote of its k neighbors. K is a positive integer, typically not very high. If "K = 1", then the object is assigned to the class of its neighbor one. In a binary context where there are only two classes, it is appropriate to choose k odd to avoid ending up in a position of equality. The K-NN classification algorithm decides the output based on the most represented class among the K neighbors. If the output is continuous, the decision to the majority does not have more sense (the values can be all different). In this case, the K neighbors' average can be assumed as the output value (Imandoust & Bolandraftar, 2013). Considering only the votes of K neighboring objects, there is a drawback due to the predominance of classes with more objects. In this case, it may be useful to weigh the neighbors' contributions to give, in the calculation of the average, greater importance according to the distance from the object considered. The choice of K depends on the characteristics of the data. Generally, as K increases, the noise that compromises the classification is reduced, but the criterion of choice for the class becomes more labile. The choice can be made through heuristic techniques (Manning & Schuetze, 1999).

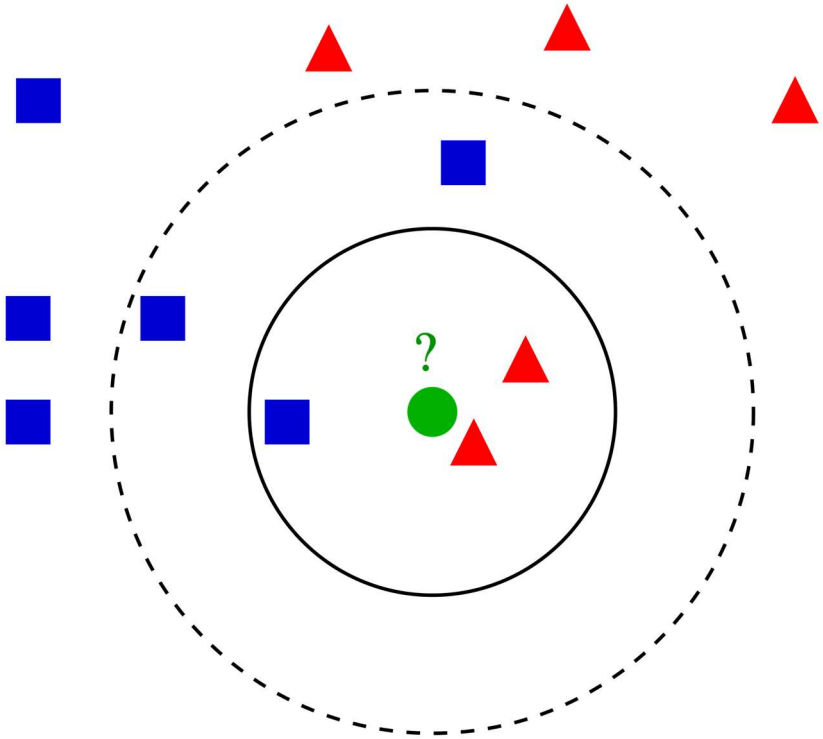

Figure 26: If k = 3 (i.e., the three closest objects are considered), then the green dot is placed in the same class as the red triangles because two triangles and 1 square are present. If k = 5, then it is placed in the same class as the blue squares because three squares and two triangles are present.

The two algorithms, NNI and K-NN, have two different purposes: the former describes the distribution of points as clustered or dispersed, the latter is a classification algorithm. There is no use in the literature of the variable K within the NNI algorithm. Given that the NNI is a continuous measure, K was integrated into the algorithm as follows: The ratio of 1) the minimum average distance of a point to its K nearest points and 2) the minimum average

distance between points if they were perfectly distributed within the area. It should be kept in mind that by using the K, the NNI can greatly exceed the expected limit value of 2.15. This latter effect is caused by the denominator of the equation, where the K is not applicable. In this sense, using a K greater than 1 can reduce the "noise" of the instrument given by the calculation of fixations (e.g., fixations that are too short and too close together).

The K-NN was computed in epochs of 1 minute for each participant with a second (K = 2) and third-order k (K = 3). One subject was excluded from the data analysis due to the low quality of recorded eye movements. Averaged K-NN values were used as the dependent variable in repeated measures ANOVA using conditions as the repeated factor. With K = 2, results showed a main effect of condition [$F_{3, 87} = 13.09$, $p < .001$] (Figure 26; Table 19). TD condition showed higher K-NN values (i.e., a more dispersed distribution of fixations) than the baseline, while in the MD condition, we obtained lower values.

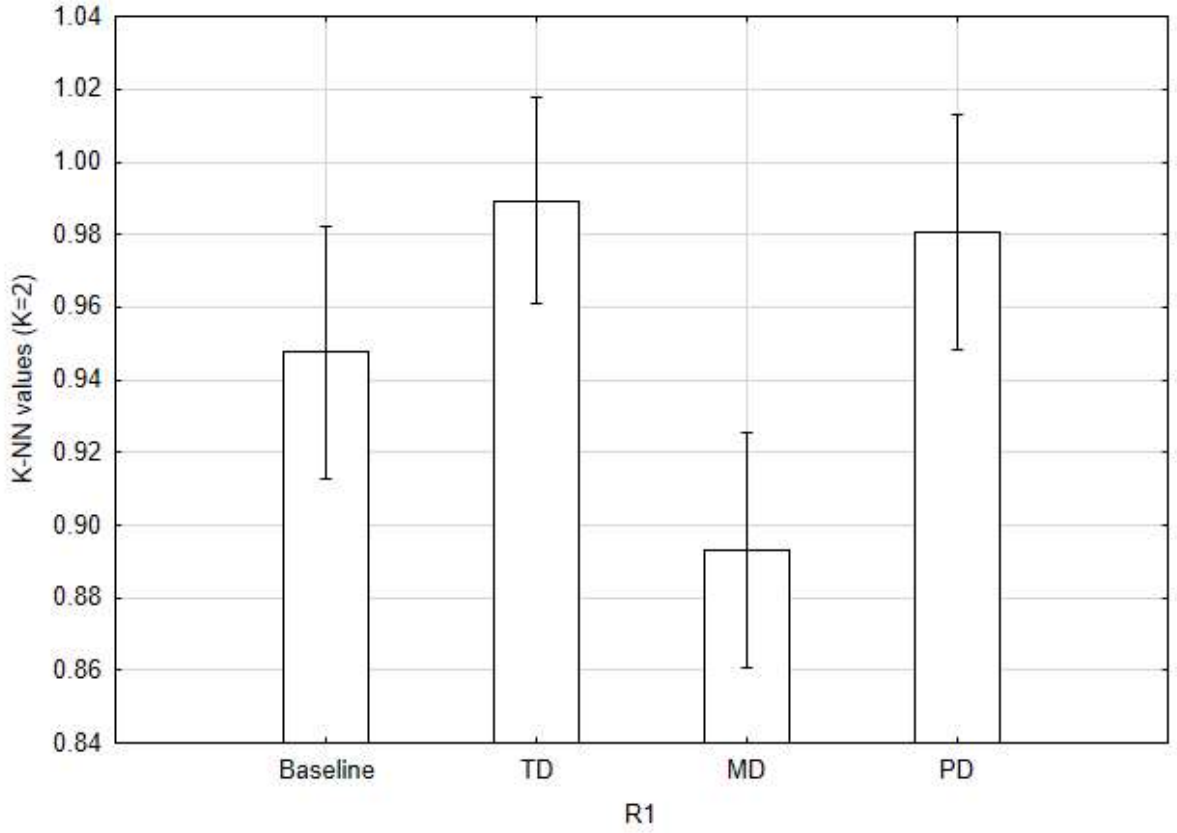

Figure 26: Average K-NN value (K=2) for the conditions compared with the baseline separately. Error bars denote .95 confidence intervals.

	TD	MD	PD
Baseline	.021	.002	.056
TD		.001	.609
MD			.001

Table 19: Post-hoc analysis carried out through the Duncan test. Pairwise comparison among K-NN (K=2) scores and conditions.

With K = 3, results showed a main effect of condition [$F_{3, 87} = 13.09$, $p < .001$] (Figure 27; Table 20). TD condition showed higher K-NN values (i.e., a more dispersed distribution of fixations) than the baseline, while in the MD condition, we obtained lower values.

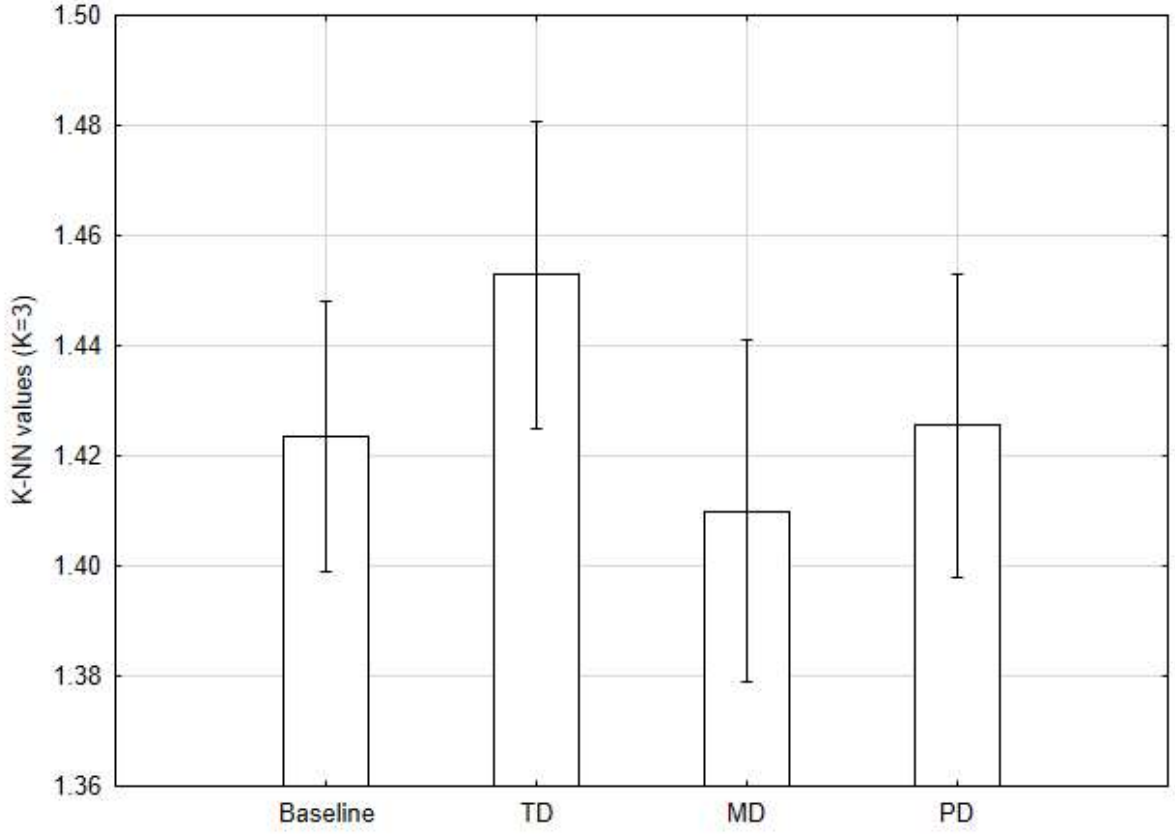

Figure 27. Average K-NN value (K=3) for the conditions compared with the baseline separately. Error bars denote .95 confidence intervals.

	TD	MD	PD
Baseline	.002	.119	.824
TD		.001	.002
MD			.092

Table 20. Post-hoc analysis carried out through the Duncan test. Pairwise comparison among K-NN (K=3) scores and conditions.

Figure 28 compares the results obtained with second-and third-order K-NN versus the classical NNI algorithm used in the first study. The plot shows 3 very similar trends, especially in relation to NNI and second-order K-NN, where the significance values are equivalent. However, in the third-order K-NN plot, a more pronounced change and loss of significance are observed than in the previous ones.

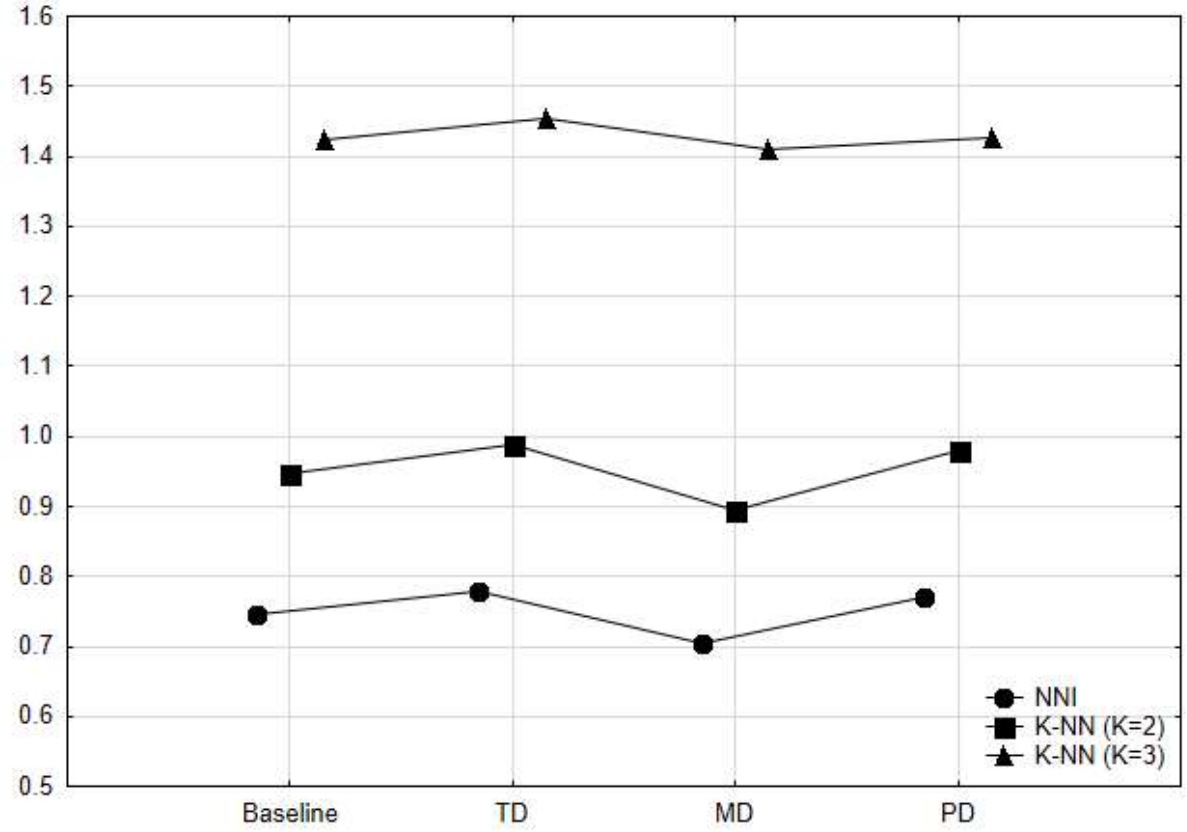

Figure 29: Comparison of values obtained by varying the K factor: NNI (K = 1) vs K = 2 vs K = 1.

Scanpath length analysis. String-edit distance calculation (Bunke, 1992; Levenshtein, 1966) is a method that could be easily and appropriately use to the scanpath. This should performed in two different stages:

- In the first, the entire visual area is split into specific areas of interest by using a grid, and then the scan path is translated into a string of characters. Each character refers to a specific AOI, and the combination provides a textual representation of the visual strategy adopted by the subject.
- In the second, the string is subjected to a sequence of transformations (inserts, deletes, and substitutions) to match a second string (for comparison).

The number of necessary transformations to match the strings allows quantifying the dissimilarity between them and, consequently, between two scan paths. This method has been adapted to compare the similarity of scanpaths (Foulsham & Kingstone, 2013; Harding & Bloj, 2010; Underwood, Foulsham, & Humphrey, 2009). Several criticisms have been raised against this measure:

- The grid is defined independently of the image content and can be too approximate in some regions or too fine in others.
- Two fixations, according to the algorithm used, can be considered distinct even if they occupy the same position.

- Defining more specific AOIs can be time-consuming and must be performed a priori. In a dynamic context, where the image changes over time, this is not always possible.

In this study, the scanpath length (the total of all saccades) was used to compare the different conditions. This allowed us to obtain a quick measure representative of the strategy used by the subject.

The saccade is considered one of the fastest movements of the human body. It varies in amplitude, duration, and maximum angular velocity (defined as peak velocity). The relationship of these three parameters has been defined as "the main sequence", indicating that the peak velocity value and the saccade duration increase systematically with the same amplitude (Di Stasi et al. 2010). The image portions where saccades are performed are typically irrelevant in the application of many searches. In addition, micro-movements recorded during a fixation, such as a tremor or rapid movements away from the focus (flicks), often count for little during high-level analysis (Salvucci, 2000). The mathematical definition of Fixation and Saccade allowed researchers to study new indexes and then observe how they vary according to mental workload. Di Stasi et al. (2016) investigated the effects of simulated flights' duration on the speed of saccadic movements made by pilots. The results show that the time spent on the task increases together with the subject's perceived fatigue, and so, the speed of the saccadic movements decreases.

The scanpath length was computed, as the sum of the amplitude of the saccade, in epochs of 1 minute for each participant. The average scanpath length values were used as the dependent variable in repeated measures ANOVA using conditions as the repeated factor. Results showed a main effect of condition ($F_{3, 87} = 14.581$, $p < .001$) (Figure 29). MD condition showed a shorter scan path, in line with NNI results. This confirms a more clustered strategy.

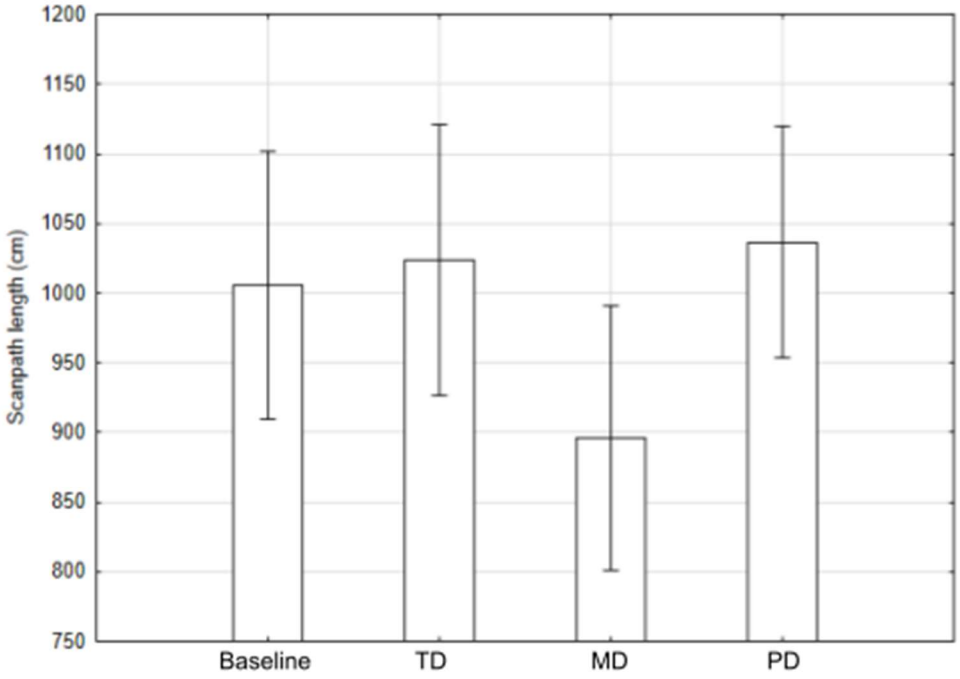

Figure 29. Average Scanpath length value for the conditions compared with the baseline separately. Error bars denote .95 confidence intervals.

Spectral analysis. The focus of this analysis was the quantitative time evolution of NNI as a task is carried out. To this purpose, data from the first study were further analyzed using spectral analysis, which is appropriately and commonly used in studying measurements collected at regularly spaced intervals of time. As described previously, in the first study, subjects performed four sessions of 10 minutes each. Therefore, A total of 40 NNI points were calculated for each subject. To obtain a more detailed plot for each condition, NNI values were recalculated using a 60-second moving window with 1-second steps. In this way, we were able to obtain a total of 542 NNI points for each condition. Figure 30 shows the average power spectra for each condition. Visual inspection of individual spectrograms showed that the four conditions provide almost identical spectral power.

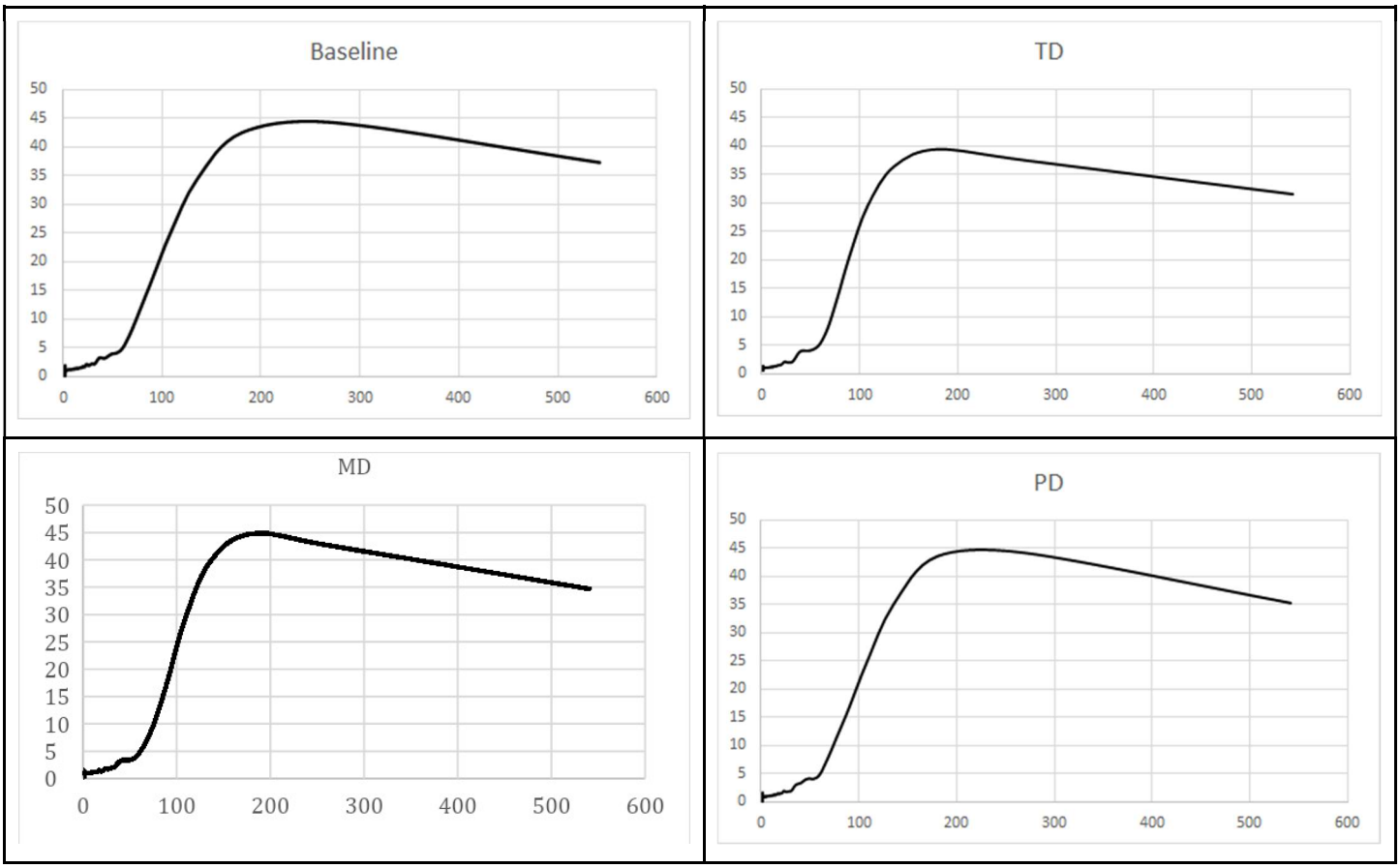

Figure 30. Power spectra representation of each condition

This result suggests that there are no observable differences in the frequency domain. However, it should be mentioned that experiments designed to study ultradian rhythms usually last for hours and make use of much longer series than those described here. The main limitation of the present account is the fact that the sessions lasted only 10 minutes.

6.1. Discussion

This latest study aimed to analyze the NNI from new perspectives, obtain new information for its interpretation concerning mental workload, and improve it by providing a more stable measure over time. The parameter K was added to the basic algorithm in the first step. The second-and third-order K-NN was calculated to obtain a more defined outcome. The results with K = 2 showed a very similar trend compared to the classical NNI. However, it should be noted that the second-order K-NN provides data relative to the minimum average distance between a point and its k nearest points. Thus, it may be assumed that this result is more filtered (less affected by data noise) than the classical minimum average distance (Distance between a point and its nearest point). This data is certainly less sensitive to outside single fixations or data considered "dirty". In some contexts, this adjustment (K=2) may have benefits. Next, we observe that as K increases, the differences between conditions tend to decrease. The choice of the parameter K is crucial to obtaining a valid result. The K should be based on the area's size, the type of task that may affect the fixation parameters, and their sampling. Validation studies will be needed to define this process better.

Regarding spectral analysis, previous studies (Di Nocera, Ranvaud & Pasquali, 2015) reported differences between low, mid, and high frequencies during flight operations. Different spectra have been observed about different flight phases, on total recordings of 38 minutes each. However, as already highlighted by the authors, usually spectral analysis is performed using a much longer time series, ranging from a few hours to several recordings. The result obtained in the latter analysis thus suggests that NNI does not vary in terms of frequency between conditions with a low or high mental workload. Future studies should be designed to specifically approach the oscillatory pattern examined here and compare it with that observed in prolonged vigilance tasks. That could be accomplished, for example, by adding a secondary reaction time task to understand whether or not the cyclic patterns in eye movements and performance data are comparable.

GENERAL DISCUSSION AND CONCLUSIONS

One of the most important challenges for research in human-technology interaction is the creation of systems capable of understanding human behavior so that automation can jump in (only) when necessary.

Indirect measures of mental workload (they all are) can be sensitive to variations in the task-load imposed on the individual. Many of them can provide only a coarse distinction between task-load levels. Others have been reported to be more fine-grained. Nonetheless, sensitivity to task-load variation is not the only important property of a successful indicator: sensitivity to different types of task demands is also important. Indeed, what we call mental workload (independently of its conceptualization) may be generated in response to changes in the task-load that, in turn, may be due to changes in the visuo-spatial component of the task (i.e., the task becomes more demanding because the individual need to look more, to find more, to discriminate more) or the task-load may be due to changes in the temporal component of the task (i.e., the task becomes faster, the interval between incoming stimuli becomes shorter, the time pressure for responding increases).

This report discusses a set of studies designed to shed light on the relationship between mental workload and ocular scanning. This topic has been covered in the Human Factors / Ergonomics literature by using different approaches, but a complete understanding of that relationship is still a long way off.

Previous studies of our laboratory have explored the opportunity to use the distribution of eye fixations as an indicator of mental workload. The Nearest Neighbor Index, a spatial statistic providing information about the distribution of points into a 2-dimensional space, was found to be sensitive to variations in mental workload. However, results obtained using the NNI were apparently different from those obtained by accredited studies using scanning randomness or "entropy" for summarizing the scanpath, therefore questioning the value of this approach. Di Nocera and Bolia (2007) have initially speculated that two processes respectively contribute to dispersion and grouping of the fixations: the temporal demand (that was manipulated in the seminal NNI studies) and the visuo-spatial demand (that was manipulated in other studies, including those featuring entropy). Here we provided empirical support for those speculations. The NNI is actually sensitive to different types of demands. The results showed high clustering when the task-load increment was obtained by changing the mental (visuo-spatial) demand and low clustering when it was obtained by changing the temporal demand. The physical demand, instead, did not affect the scanpath, possibly because our manipulation of this dimension was not appropriate or because the ocular behavior is not sensitive to the manipulation of the physical component. Albeit results showed a significant increase in the subjective estimates of physical demand, the effect did not extend to the overall workload ratings nor to the analysis of the scanpath. Likely, the Tetris game involved minimal physical effort, and the manipulation was not effective. To overcome this limitation, future studies could consider several options. One potential solution could be to manipulate the game controls, producing frequent keypress failures in the high task-load condition. Alternatively, the keypress force could be manipulated in the high task-load condition to make the task more effortful.

In one of the studies reported here, the NNI was directly compared to the entropy approach. Results showed an overall increase of difficulty after the first few minutes of performance that reflected in both measures of mental workload. After two minutes, the search task generated both a stereotyped dwell pattern (consistent with the entropy prediction) and fixations grouping (consistent with the fixations distribution prediction). In other words, the two indices were found to be both sensitive to changes in the visuo-spatial demand, and the plots were highly overlapping. Such a result sorts out the issue of the differences found between the two indicators, showing how that exclusively depends on the type of demand imposed. Also, results demonstrated that a dispersed fixation pattern (or moderately grouped) is not equivalent to high randomness in visual exploration. The two scanpath analysis algorithms show the same trend. From the post-hoc analyses, a cut-off of the values is observed starting in the fourth minute. However, compared to the performance and mental workload data, the cut-off occurs only after the sixth minute of activity. This difference suggests that the change in visual exploration strategy anticipated the decline in-game performance on the fourth minute. Further studies should be conducted to confirm this effect. In high-risk settings, the anticipation of critical events and the operator's mental state is essential.

A third study was run 1) to determine whether the scanpath analysis via NNI could be used as a trigger for adaptive automation and 2) to identify the optimal level of automation to be applied. A Tetris "autopilot", able to take total control of the game, was designed as the best solution to avoid game-over in critical situations. Results of the analysis carried out on the ocular metric showed a significant interaction effect between the difficulty and automation, suggesting that automation was effective in helping the subject in terms of performance and NNI.

We further addressed the use of the NNI as a trigger for adaptive automation in a study aimed at observing what happens, in terms of ocular strategies, before and after the present/absence of automation. The presence or absence of automation support was classified into four categories: Provided when needed **(NP)**; Provided when not needed **(nNP)**; Not provided when needed **(NA)**; Not provided when not needed **(nNA)**. Unfortunately, results showed high variability, probably caused by the low number of trials and the high dynamism of the task, where a single error can lead to game-over. However, it should be noted that this experiment was based on the calculation of a 10-minute baseline that allowed us to calculate threshold values. Future studies should consider a more durable baseline to provide more accurate values for each subject.

The last part of this report was devoted to exploring new methods for analyzing the ocular pattern recorded in the first experiment. The analyses compared NNI scores with those measured by first- and second-order K-NN. The results with K = 2 showed a very similar trend compared to the classical NNI. Instead, with K = 3, the differences between conditions tend to decrease. The choice of the parameter K is crucial to obtaining a valid and indicative result. This should be based on the area's size, the type of task that may affect the fixation parameters, and their sampling. Subsequently, NNI points (provided by 60s moving windows with the step of 1s) were used in a spectral analysis. The result suggests that NNI does not vary in terms of frequency between conditions with a low or high mental workload. However, as already highlighted in a previous study (Di Nocera, Ranvaud & Pasquali, 2015), usually spectral

analysis is performed using much longer time series ranging from a few hours to several days of recordings.

In conclusion, the NNI seems to be a good indicator of mental workload, but only under specific conditions:

1) First, it is necessary to estimate a single operator's baseline while performing a specific task. It should be considered that the interfaces have different visual characteristics. For example, during a driving task, the driver has to watch the fuel level, navigation system, mirrors, and road. These elements require their attention, and, therefore, the generated fixations will be allocated to specific visual areas pertaining to that task. The baseline aims to identify the visual exploration strategy used during an optimal condition of the task (neither too easy nor too hard) and the subject's psychophysical state.
2) Subsequently, the scanpath is monitored and processed in real-time (minute by minute) using the Nearest Neighbor Index. The values thus obtained are compared with the average value provided by the baseline. Considering the fourth study described here, the NNI values analyzed separately fluctuate within a wide range, causing a large variability without the possibility to distinguish the different conditions. However, when we analyze the average of 10 NNI values (provided by 10-minute sessions), a more stable result allows us to distinguish the difficult conditions (compared to the baseline) and the type of task demand imposed. In a realistic context, using more values generated every 5 or 10 minutes would provide more accurate and stable data, less affected by sampling errors or extraneous visual elements. Supervision and control tasks that may last several hours are common in many high-risk contexts. The NNI should be able to monitor the level of vigilance and mental workload perceived by the operator. To accomplish this, a psychophysiological measure generated every 10 minutes should be sufficient to reduce the risk of incidents in these contests.
3) The last recommendation concerns the proper level of automation to be applied when an anomaly in the NNI plot is detected. In the Tetris used here, the right automation would be slowing down the piece's descent (using the same speed as the easy level as set out in studies 1, 3, and 4). However, this system does not seem realistic outside of a laboratory setting. In addition, any automation changes the nature of the task and risks invalidating the previously calculated baseline. A possible solution is to calculate a second baseline to be used when the automation is active. Future studies should consider these aspects to use the NNI as a trigger in an adaptive automation system.

REFERENCES

Bunke, H. (1992). Recent advances in string matching. Advances in structural and syntactic pattern recognition, 3-21.

Camilli, M., Terenzi, M., & Di Nocera, F. (2007). Concurrent validity of an ocular measure of mental workload. In D. de Waard, G.R.J. Hockey, P. Nickel, and K.A. Brookhuis (Eds.), Human Factors Issues in Complex System Performance (pp. 117-129). Maastricht, the Netherlands: Shaker Publishing.

Camilli, M., Terenzi, M., & Di Nocera, F. (2008). Effects of temporal and spatial demands on the distribution of eye fixations. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 52(18), 1248-1251."

Clark, PJ, & Evans, RC, 1954 "Distance to nearest neighbor as a measure of spatial relationships in populations" Ecology 35 445-453

Di Nocera, F., & Bolia, R. S. (2007). PERT networks as a method for analyzing the visual scanning strategies of aircraft pilots. Proceedings of the 14th International Symposium on Aviation Psychology (pp. 165–169). Dayton, OH: Wright State University.

Di Nocera, F., Camilli, M., & Terenzi, M. (2007) A random glance at the flight deck: Pilots' scanning strategies and the real-time assessment of mental workload. Journal of Cognitive Engineering and Decision Making, 1(3), 271-285.

Di Nocera, F., Ranvaud, R., & Pasquali, V. (2015) Spatial pattern of eye fixations and evidence of ultradian rhythms in aircraft pilots. Aerospace Medicine and Human Performance, 86(7), 647-651.

Di Nocera, F., Ricciardi, O., Mastrangelo, S., Torres, E., Bordignon, M., Marcolin, F. (2020). Scanpath analysis into the wild: the spatiotemporal distribution of fixations as an indicator of driver's mental workload. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 64(1):371-375.

Di Stasi, L. L., Renner, R., Staehr, P., Helmert, J. R., Velichkovsky, B. M., Cañas, J. J., ... & Pannasch, S. (2010). Saccadic peak velocity sensitivity to variations in mental workload. Aviation, Space, and Environmental Medicine, 81(4), 413-417.

Dillard, M. B., Warm, J. S., Funke, G. J., Funke, M. E., Finomore, V. S., Matthews, G., Parasuraman, R. (2014) The Sustained Attention to Response Task (SART) Does Not Promote Mindlessness During Vigilance Performance. Human Factors, 56(8), 1364–1379.

Fidopiastis, C. M., Drexler, J., Barber, D., Cosenzo, K., Barnes, M., Chen, J. Y., & Nicholson, D. (2009, July) Impact of automation and task-load on unmanned system operator's eye movement patterns. In International Conference on Foundations of Augmented Cognition, pp. 229-238, Springer, Berlin, Heidelberg.

Foy, H. J., & Chapman, P. (2018). Mental workload is reflected in driver behaviour, physiology, eye movements and prefrontal cortex activation. Applied ergonomics, 73, 90-99.

Foulsham, T., & Kingstone, A. (2013). Fixation-dependent memory for natural scenes: an experimental test of scanpath theory. Journal of Experimental Psychology: General, 142(1), 41.

Harding, G., & Bloj, M. (2010). Real and predicted influence of image manipulations on eye movements during scene recognition. Journal of Vision, 10(2), 8-8.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), Advances in psychology, 52. Human mental workload 139–183. North-Holland. https://doi.org/10.1016/S0166-4115(08)62386-9

Imandoust, S. B., & Bolandraftar, M. (2013). Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background. International Journal of Engineering Research and Applications, 3(5), 605-610.

Levenshtein, V., (1966) Binary codes capable of correcting deletions, insertions and reversals, Soviet Physics Doklady, 10(8): 707-710.

Manning, C.D & Schuetze, H. (1999). Foundations of Statistical Natural Language Processing.The MIT Press

Marshall, S. P. (2007). Identifying cognitive state from eye metrics. Aviation, space, and environmental medicine, 78(5), B165-B175.

Neumann, D. L., & Lipp, O. V. (2002). Spontaneous and reflexive eye activity measures of mental workload. Australian Journal of Psychology, 54, 174–179.

Salvucci, D. D., & Goldberg, J. H. (2000, November). Identifying fixations and saccades in eye-tracking protocols. In Proceedings of the 2000 symposium on Eye tracking research & applications. 71-78

Smith, C. E. (1998). Modeling high sinuosity meanders in a small flume. Geomorphology, 25(1-2), 19-30.

Stark, B. & Young, D. (1981). Linear Nearest neighbor Analysis. American Antiquity. 46. 284. 10.2307/280209.

Tattersall, A. J., & Foord, P. S. (1996) An experimental evaluation of instantaneous self-assessment as a measure of workload. Ergonomics, 39(5), 740-748.

Tole, J. R., Stephens, A. T., Vivaudou, M., Ephrath, A. R., & Young, L. R. (1983) Visual scanning behaviour and pilot workload.

Trimmel, M., & Huber, R. (1998). After-effects of human-computer interaction indicated by P300 of the event-related brain potential. Ergonomics, 41(5), 649-655.

Underwood, G., Foulsham, T., & Humphrey, K. (2009). Saliency and scan patterns in the inspection of real-world scenes: Eye movements during encoding and recognition. Visual Cognition, 17(6-7), 812-834.

Wierwille, W. W., & Eggemeier, F. T. (1993). Recommendations for mental workload measurement in a test and evaluation environment. Human factors, 35(2), 263-281.