applied sciences

MDPI

*Article*

# Studying the Role of Visuospatial Attention in the Multi-Attribute Task Battery II

**Daniel Gugerell [1,*], Benedikt Gollan [2], Moritz Stolte [1] and Ulrich Ansorge [1,3,4]**

[1] Faculty of Psychology, University of Vienna, 1010 Vienna, Austria; moritz.stolte@univie.ac.at (M.S.); ulrich.ansorge@univie.ac.at (U.A.)
[2] Research Studios Austria, 1090 Vienna, Austria; benedikt.gollan@researchstudio.at
[3] Vienna Cognitive Science Hub, University of Vienna, Vienna, Austria
[4] Research Platform Mediatised Lifeworlds, University of Vienna, Vienna, Austria
* Correspondence: daniel.gugerell@univie.ac.at

**Abstract:** Task batteries mimicking user tasks are of high heuristic value. Supposedly, they measure individual human aptitude regarding the task in question. However, less is often known about the underlying mechanisms or functions that account for task performance in such complex batteries. This is also true of the Multi-Attribute Task Battery (MATB-II). The MATB-II is a computer display task. It aims to measure human control operations on a flight console. Using the MATB-II and a visual-search task measure of spatial attention, we tested if capture of spatial attention in a bottom-up or top-down way predicted performance in the MATB-II. This is important to understand for questions such as how to implement warning signals on visual displays in human–computer interaction and for what to practice during training of operating with such displays. To measure visuospatial attention, we used both classical task-performance measures (i.e., reaction times and accuracy) as well as novel unobtrusive real-time pupillometry. The latter was done as pupil size covaries with task demands. A large number of analyses showed that: (1) Top-down attention measured before and after the MATB-II was positively correlated. (2) Test-retest reliability was also given for bottom-up attention, but to a smaller degree. As expected, the two spatial attention measures were also negatively correlated with one another. However, (3) neither of the visuospatial attention measures was significantly correlated with overall MATB-II performance, nor with (4) any of the MATB-II subtask performance measures. The latter was true even if the subtask required visuospatial attention (as in the system monitoring task of the MATB-II). (5) Neither did pupillometry predict MATB-II performance, nor performance in any of the MATB-II's subtasks. Yet, (6) pupil size discriminated between different stages of subtask performance in system monitoring. This finding indicated that temporal segregation of pupil size measures is necessary for their correct interpretation, and that caution is advised regarding average pupil-size measures of task demands across tasks and time points within tasks. Finally, we observed surprising effects of workload (or cognitive load) manipulation on MATB-II performance itself, namely, better performance under high- rather than low-workload conditions. The latter findings imply that the MATB-II itself poses a number of questions about its underlying rationale, besides allowing occasional usage in more applied research.

**Keywords:** pupil dilation; eye-tracking; MATB-II; task demands; attention capture



**Citation:** Gugerell, D.; Gollan, B.; Stolte, M.; Ansorge, U. Studying the Role of Visuospatial Attention in the Multi-Attribute Task Battery II. *Appl. Sci.* **2024**, *14*, 3158. https://doi.org/10.3390/app14083158

Academic Editors: Zbigniew Gomolka, Damian Kordos, Ewa Dudek-Dyduch and Bogusław Twaróg

Received: 10 January 2024
Revised: 28 March 2024
Accepted: 2 April 2024
Published: 9 April 2024



## 1. Introduction

### *1.1. Impact Statement*

Human cognitive aptitude for specific applied tasks is often measured with tests or task batteries. These batteries mimic important surface characteristics of the applied task in question. As an example, take the Multi-Attribute Task Battery II (MATB-II). The MATB-II presents components of a flight console on a computer screen, and participants are requested to perform a sequence of subtasks typical for flying (e.g., monitoring for display

changes or tracking of moving objects). Batteries such as the MATB-II have potentially high ecological or external validity [1,2]. This is due to their real-world similarity. However, often less is known about their internal validity: it is unclear what underlying human function or mechanism they are measuring because a theoretical model connecting basic human functioning with task performance is lacking. Yet, this is important: to understand the basic functions involved helps to improve displays and training. For example, display signals can be designed to fit to the processing capabilities of the human user. Where this is not possible, known difficulties (e.g., high demands imposed by a faint or peripheral visual signal) could at least be made easier to process for humans by systematic training (e.g., practicing the systematic scanning of the display for the nonsalient signals).

Regarding external versus internal validity, the situation is typically the opposite in a controlled experimental laboratory task. These controlled tasks often measure specific well-described underlying functions and are, therefore, of high internal validity. However, these controlled laboratory tasks are less similar to real-world tasks. Thus, their external and ecological validity is doubtful. It is unclear if they could be used successfully to discriminate individual aptitude in a specific real-world task. For example, the effects of separate cognitive functions or mechanisms (e.g., of memory capacity and inhibition of interference) might not simply add up in a complex real-world task.

In the current study, we, therefore, aimed to understand the connection between performance in MATB-II and one specific experimental task measuring visuospatial attention. Visuospatial attention denotes the selection of locations in the visual environment. This ability is highly relevant for flight performance. For example, visuospatial attention is used to monitor or track different parts of a flight console. However, do individual capabilities of visuospatial attention predict performance in the MATB-II? Or is MATB-II performance driven by other cognitive factors such as memory or the ability to switch between tasks? In addition, maybe humans steer their visuospatial attention in the MATB-II entirely differently from how they do it in a typical visual search task used to measure visuospatial attention under controlled laboratory conditions.

In the current study, in line with the latter possibilities, we did not find any significant correlations between human performance on an experimental task measuring visuospatial attention and the MATB-II. This was also the case for unobtrusive measures of cognitive performance. In summary, thus, the present work was concerned with the internal validity of the MATB-II. Questions regarding the external validity or ecological validity of both the MATB-II and of our controlled measure of visuospatial attention were not addressed. However, the fact that we did not test external validity—for example, that we did not test the sensitivity of the MATB-II for the correct discrimination between experts (pilots) and novices—might have created a caveat. Because we used an opportunity sample consisting of mostly students and a few professional pilots, cognitive performance in this sample might have simply created too little variance (though we varied task difficulty on the MATB-II to create some variance in cognitive performance). Of lesser importance, in the current study, we found interesting and unexpected effects in MATB-II performance as a function of task: performance in the MATB-II increased rather than decreased with increasing task demands (here, the number of subtasks to be performed per unit of time; see Supplement Tables S2 and S3), and we observed that unobtrusive pupillometry could be used to discriminate between stages of subtask performance.

### *1.2. Theoretical Background*

The diagnosis of individual cognitive aptitudes such as that of selective visuospatial attention or working memory capacity can take on different forms, from test batteries, over real-world tasks, to relatively pure tests of individual cognitive functions in controlled experiments [3–7]. In the domain of visuospatial attention, Weichselbaum et al. [8] have recently argued for the usage of experimental tasks because they fulfill two important criteria of diagnostic measurements: They are of high internal validity because they measure different forms of capture of visuospatial attention—top-down dependent capture of

visuospatial attention based on current search goals versus bottom-up capture of attention due to high local feature contrasts or salience—in a relatively pure way; that is, free from confounding or contaminating sources of variance such as task shifting, working memory demands, or modality changes. These experimental task measures also showed test-retest reliability in the form of significant correlations between measures of attention across different measurement time points [8,9].

Here, we used the experimental measurement of visuospatial attention to see if relatively pure measures of visuospatial attention can help us to understand performance in an applied test: the Multi-Attribute Task Battery (II, MATB-II) [10]. The MATB-II uses a simplified version of a flight console. The MATB-II can be presented on a computer monitor, and participants must work on different tasks related to the control of the simulated flight console. For their MATB-II performance, participants, thus, have to switch between tasks, with different tasks associated with different areas of the console and with tasks presented in an unforeseeable sequence. For instance, participants have to monitor two buttons in the upper-left hand corner for color changes (from green to gray and from gray to red) and have to respond to the color changes by left-clicking with the mouse on the corresponding changed buttons (system monitoring task). This task is embedded in a sequence of other tasks such as the tracking of a moving crosshair (tracking task) presented to the right of the light and scales, or responses in the lower left of the display to audio presented via headphones (communication task). An in-depth explanation and description of the MATB-II can be found in Section 2.4 below.

Importantly, we chose the MATB-II for an investigation with our experimental visuospatial attention measurement because especially the system monitoring task of salient color changes in the upper left corner during MATB-II performance could benefit either from bottom-up capture of visuospatial attention by salient stimuli, here, color changes, or from top-down control of visuospatial attention, here, to look for particular colors (e.g., the color red) [11–15]. To tell if bottom-up or top-down attention is responsible for MATB-II performance is important. If salient features capture attention in a stimulus-driven or bottom-up fashion, pilots would not need any training and participants would not need any instructions regarding what color changes to look for. A salient color change would capture attention, ensuring that pilots or participants pay attention to the relevant signal. However, the situation is different if top-down control accounts for what pilots and participants pay attention to. Here is an example: Suppose you were to pick up a friend in front of a crowded station. To successfully spot your friend in the crowd, you would have to know what he or she looks like and search for him or her by these known features. In an analogue case during the control of the flight console, pilots and participants would need proper training and instructions to search for and find a critical signal on the console or display. In this case, it would probably also be helpful to use fewer different relevant features to search for across different subtasks in the MATB-II or in flight-console operation, so as not to hinder switching between tasks should a signal onset happen to indicate a necessary task switch (e.g., a blinking lamp to use fuel from a specific pump, see below for details).

In the present study, we correlated performance in the MATB-II with experimental measures of bottom-up versus top-down capture of visuospatial attention. In this way, we were hoping to understand if one of these types of visuospatial attention capture contributes to the performance in the MATB-II in general, and in the system monitoring task of the MATB-II in particular. We present the correlations in Section 3.2.

We also took care to create sufficient performance variance within the MATB-II. This was necessary for the planned correlation or regression analyses because these analyses depend on sufficient variance. To achieve sufficient variance, we varied task demands or cognitive load in the MATB-II in two steps. Specifically, we varied the frequency of tasks and task shifts per unit of time in the MATB-II. Here, a higher frequency of tasks and task shifts per unit of time corresponded to a higher workload. In contrast, a lower frequency of tasks and task shifts per unit of time corresponded to a lower workload.

Obviously, visuospatial attention—the selection of some locations on the console while ignoring others—is involved in MATB-II task performance. For example, one can either track an object with the eyes or look at a changing button for the system monitoring task, but one cannot select both objects with the eyes at the same time. Thus, we would have expected that visuospatial attention and MATB-II task performance are significantly correlated. However, for at least two reasons, we might have failed to find a significant correlation. First, it is not clear if capture of visuospatial attention is the domineering factor for MATB-II task performance. For example, performance on the MATB-II also requires task shifting, and task-shifting abilities that are relatively independent of visuospatial attention could be more decisive than visuospatial attention for overall performance on the MATB-II. Secondly, related to this point, in more applied and real-world tasks such as the MATB-II, the control of visuospatial attention can take on forms that are relatively distinct from the types of capture of visuospatial attention that account for typical visual-search task performance. For example, in many applied and real-world tasks, humans know where things are located and can, thus, systematically shift their attention to a specific location in anticipation of an expected task at this location: if humans expect a task, they could shift attention to a specific location (e.g., to the area of the console at which a cross-hair needs to be tracked). In contrast, no anticipatory attention shifts to a particular location are possible in our experimental visual-search task because, from trial to trial, a to-be-searched-for visual target is equally likely to appear at any possible stimulus location.

In the current study, in addition to the traditional performance measures such as reaction times and numbers of errors, we also looked at pupillary responses. These are known to increase in response to increased task demands or cognitive load [16–22]. For example, Ahlstrom and Friedman-Berg (2016) found that average pupil size increased when controllers used a static storm forecast tool compared to when controllers used a dynamic forecast tool [23]. For the unobtrusive measurement of the pupillary responses, we employed a cognitive load algorithm that automatically models and subtracts size changes due to the pupillary light response based on empirical models of the pupillary light response and camera-based brightness measures, thereby, providing a measure of cognitive load free of this source of pupil size variation [22,24–26]. Here, we did not find any effects of our manipulation of task performance (see Section 3.3). However, when we looked for causes for the lack of an average effect of our load manipulation on pupillary responses, we found that different stages of on-task performance were not all equally correlated with a pupillary response, such that on average, pupillary responses to task-load manipulations might have been washed out (see Section 3.4).

One should also note that the pupil size is related to factors other than workload (or task demands) and light, such as emotions, arousal, memory content, or pharmacological agents [27–29]. We had no reason to suspect that the one or the other of these influences was systematically confounded with the steps of our manipulations. Thus, we did not control for the corresponding influences. However, the broad variety of influences implies the relatively unspecific nature of the pupillary response, meaning that there are also disadvantages to the method besides its advantages (e.g., its unobtrusiveness), such as a relatively high level of noise brought about by the different influences.

## 2. Materials and Methods

### 2.1. Participants

The study included 53 participants; 5 were trained pilots (all male), and 49 were psychology students from the University of Vienna (30 female, 19 male). Originally, we intended to include more trained pilots. However, COVID-19 restrictions applied at the time of data acquisition, preventing a larger sample size of trained pilots. The sample has, thus, been gathered opportunistically. It is, therefore, not representative of a wider population. Sample size was based on an a priori calculation assuming a large effect size ($\eta^2 = 0.20$), aiming for a statistical power of 0.8, and allowing an alpha error of 0.05. Regarding the MATB-II, participants were randomly assigned to either the low-workload group

(26 participants, three of which were pilots) or the high-workload group (27 participants, two of which were pilots). Students received partial course credit. Pilots received a small monetary compensation for their time. Prior to the experiment, informed consent was obtained from all participants. Ethical approval (No EK_00644) was obtained from University of Vienna's Ethical Review Board.

### *2.2. Apparatus*

The experiment was conducted in a dimly lit room, where the only relevant light-source was the monitor. The tasks were presented via a 31 cm × 28.5 cm monitor (resolution, 1920 pixels × 1080 pixels; 60 Hz screen refresh rate). Participants were seated in front of the monitor, with their gaze straight ahead, centered on the screen. Viewing direction and distance (60 cm) were supported by a chinrest. Participants wore a mobile, video-based eye-tracker (Pupil Labs, Berlin, Germany; sampling at 60 Hz, with an estimated gaze accuracy of 0.6° (according to manufacturer)). A PC running Windows 10 (Microsoft, Redmond, WA, USA) and Pupil Labs software version 3.4.0 (pupil-labs.com/pupil/, accessed on 9 September 2021) was connected to the eye-tracker for recording of pupil size, eye movements, and the external visual surroundings. A picture of the setup can be seen in Figure 1. This picture only serves illustrative purposes. During testing, the eye-tracker was connected to the computer (not depicted), and the lights were dimmed (which is not the case in this figure).

**Figure 1.** Experimental setup.

Participants also wore stereo headphones (RP-HT265, Panasonic) operating with a standard volume of 25 on a Windows 10 computer.

### *2.3. Eye-Tracking Data Processing*

The pupil data were exported using Pupil Player v3.0.7, with a minimum data confidence of 0.6. Confidence is an assessment by the pupil detector on how sure it is about the measurement. This measurement is taken for each frame and each eye. Pupil Labs suggests that any confidence value exceeding 0.6 is useful data. Further data processing was done using PyCharm Community Edition 2020.2.3. The whole dataset was reduced to a single eye, which was chosen by its higher overall average confidence. The algorithm used a

3D-corrected pupil size measure that takes looking direction into account [25]. To diminish effects of uncontrolled light sources, the overhead lighting in the room was switched off during testing, leaving the monitor in front of the participant as the only light source during testing. The algorithm for the extraction of a pupillary load index used the brightness measured by the scene- or world-camera located over the right eye. The measured brightness was used to model a pupillary light response on an individual basis, separately for each participant. No further weighting of the average measured lightness by the camera (e.g., a stronger weighting of parts of the scene) was applied. The modelled light-elicited response of the pupil was subtracted to arrive at a load measure of the pupillary response.

### *2.4. Procedure and Task*

The experiment consisted of three blocks. The first and third block of the experiment were experimental measurements of capture of visuospatial attention in visual search tasks in which participants had to search for a color-defined target and report the orientation of a cross inside the target (upright or oblique). A singleton distractor was sometimes presented together with the target and away from the target. The singleton distractor could be of a relatively target-similar color or of a relatively dissimilar color, and it was expected to interfere with target search. These blocks consisted of three rounds of 84 trials each. The visual-search task was split into two blocks, one before and one after the second block, so that we could calculate a test-retest reliability of the measures of visuospatial attention. The second block, in-between the visual-search blocks, was the Multi-Attribute Task Battery (MATB-II). Before the experiment started and after each block, the eye-tracker was calibrated for each participant. To ensure proper calibration throughout the experiment, a fixation check was executed at the start of each trial in Blocks 1 and 3. Figure 2 shows an example of a trial used in the visual search task, adapted from Weichselbaum and Ansorge (2018) [9].

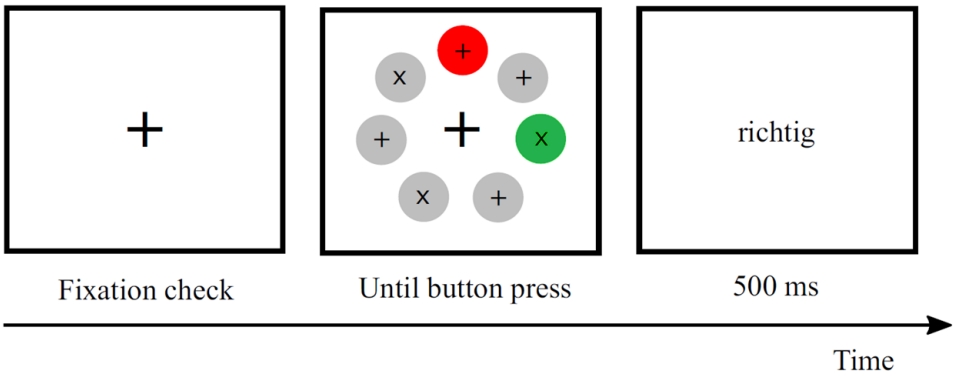


**Figure 2.** Example of a trial in the visual search task (Blocks 1 and 3).

Before the first block started, participants were shown the target color (below in coordinates of the L*a*b* system; L*: luminance; a*: red/green value; b*: blue/yellow value), which was fixed throughout the experiment and either Red 1 (L*a*b* = 62.7/79.0/65.7), Green 1 (L*a*b* = 62.5/−69.4/52.5), Red 2 (L*a*b* = 62.0/76.7/21.1), or Green 2 (L*a*b* = 62.3/−15.8/52.7), with color balanced across participants. Participants also received an explanation of what the trials would look like, and that their reaction time, as well as the number of correct responses mattered. Each trial started with the presentation of a black fixation cross at screen center for 500 ms. Here, we conducted a fixation check. If necessary, we conducted a drift correction of the eye tracker during the fixation check. Next, the target display was presented. It consisted of seven discs. One disk was in the participant's target color. The other six were either all color-nonsingletons in a neutral, gray color (L*a*b* = 62.0/12.7/−35.8)—these were the distractor-absent trials—or five gray nonsingletons plus one singleton-color distractor—these were the distractor-present trials. Each participant saw two types of distractor-present trials: trials with

a target-similar or (top-down) matching singleton-color distractor (e.g., if Green 1 was the target color, the target-similar distractor would have been presented in Green 2); and trials with a target-dissimilar or nonmatching singleton-color distractor (e.g., if Green 1 was the target color, the target-dissimilar distractor would have been presented in Red 1). Distractor-absent, distractor-present/target-similar, and distractor-present/target-dissimilar trials were presented separated in rounds of the first and the last block—that is, those conditions were the same for the 84 trials (e.g., 7 possible target positions × 6 possible distractor positions × 2) of each round (e.g., 84 distractor-absent trials before 84 distractor-present/target-similar trials before 84 distractor-present/target-dissimilar trials). By blocking distractor conditions, we decreased the likelihood of shifts between different top-down search settings even further. This was done to ensure that visuospatial attention rather than task shifting explained visual-search task performance. However, by these runs of different conditions in the visual-search task blocks, we might have also inadvertently encouraged suppression of attention capture, especially by the more target-dissimilar, nonmatching distractor: it is known that repeating distractor colors might help establish proactive suppression of the misleading distractor [30–32]. Throughout the search task, participants had to search for the disk in their instructed target color and respond by clicking either the left or right mouse-button, depending on whether the symbol inside the target disc was a "+" or an "×". After the button press, a feedback display was shown, telling the participants whether they clicked the correct (German word "richtig") or the incorrect (German word "falsch") button. For each participant, there was one target color, a particular order in which the rounds of the distractor-absent, distractor-present/target-similar, and distractor-present/target-dissimilar conditions were realized, and a specific stimulus-to-response mapping (i.e., whether an "×" required the left and the "+" required the right button press, or vice versa). To ensure that participants understood the task, they practiced the search task for at least 20 trials in the distractor-absent condition before data collection started. The task was practiced in the distractor-absent condition, as this was probably the easiest condition.

In the second block, participants were introduced to the MATB-II (see Figure 3). The MATB-II is a computer-based task battery designed to evaluate operator performance and workload with a simplified simulated cockpit console. It requires the simultaneous performance on and unforeseeable switches between several subtasks: system monitoring, tracking, communication, and dynamic resource management tasks (https://matb.larc.nasa.gov/?doing_wp_cron=1649083621.1180279254913330078125, accessed on 3 August 2021). To perform the tasks, participants used a joystick (Model: Logitech Attack 3) and a standard computer mouse.

The colored rectangles in Figure 3 around the task-specific locations were not shown to the participants. They are just helpful illustrations for referencing particular locations. On the upper left, the region surrounded by a red rectangle, numbered 1, as well as the region right below it, surrounded by a blue rectangle with the number 2, is the "System Monitoring" task. The area surrounded by the green rectangle, numbered 3, is the "Tracking" task. Bottom left, the area within the yellow rectangle and the number 4 is the "Communication" task, and to its right, the pink rectangle, numbered 5, is the "Resource Management" task. All tasks will be further explained in the following paragraphs.

System Monitoring: This task requires participants to monitor two warning lights ("F5" and "F6") in the areas within the red boundaries of Figure 3, designated by the number 1. Participants have to monitor that "F5" stays on/green and that "F6" stays out/gray. If either of those states changes ("F5" turns out/gray or "F6" turns on/red), participants must respond by left-clicking the corresponding display button. Additionally, the dark-blue pointers in the scales "F1" to "F4" in the areas of Figure 3, surrounded by blue outlines and designated by the number 2, have to stay within a certain range around

the midpoint. If a pointer deviates too much from the midpoint, the participant has to correct/reset its position to the midpoint by left-clicking on the corresponding scale.

Tracking: participants are asked to use a joystick to track the moving circle with the crosshair as a cursor in the area with the green boundary in Figure 3, designated by the number 3.

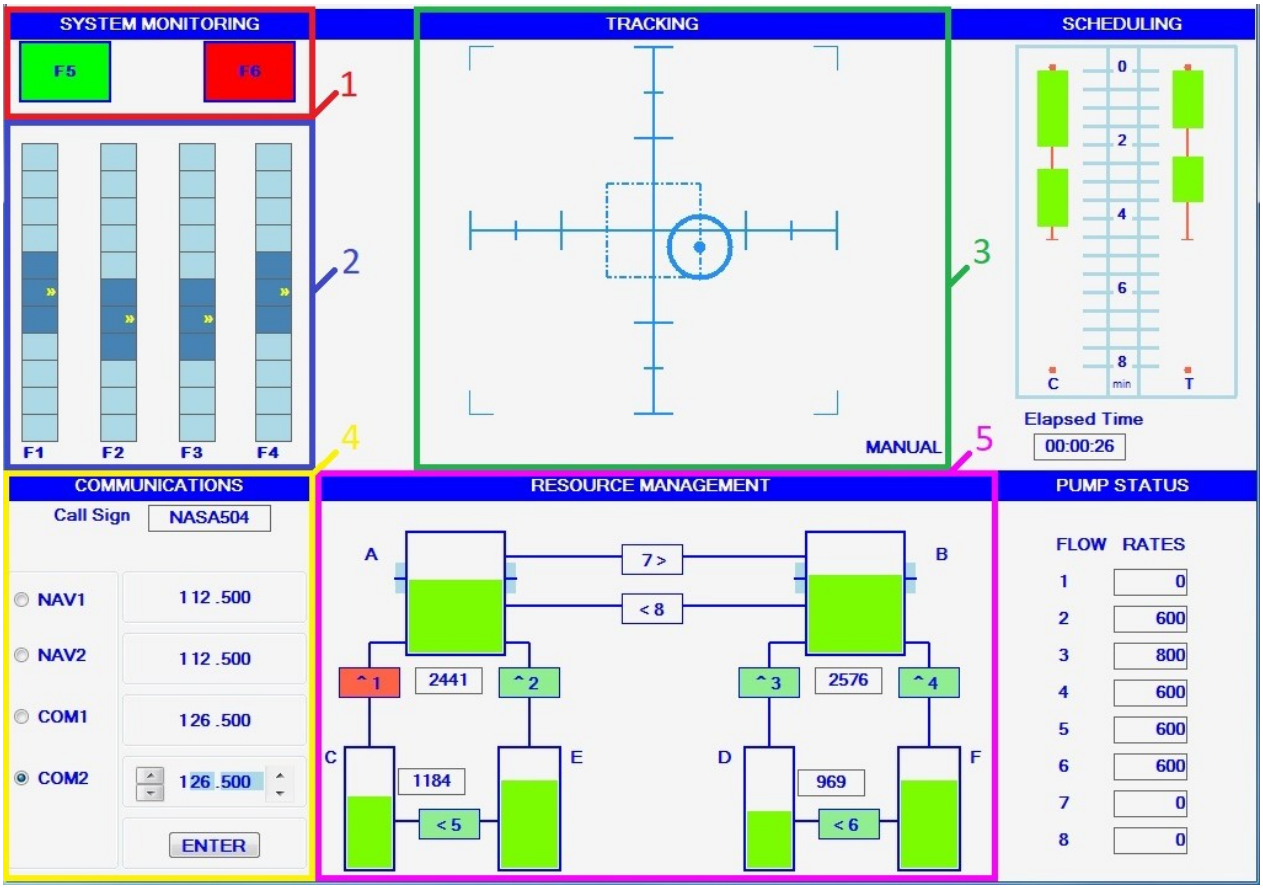


**Figure 3.** Multi-Attribute Task Battery (MATB-II).

To note, the "Scheduling" timeline to the right of this area is also related to tracking. It shows when the tracking task has to be performed manually and when it switches to "autopilot". The two green bars on the right, above the small "T" (=Tracking), signal to the participant in advance that the tracking task must be performed manually. As soon as the green bar vanishes, leaving behind only the thin orange line, the tracking task was taken over by the autopilot, giving the participants some time to focus on other tasks. The scale in the middle of this sector (located between the "C" and the "T") serves as a timeline in minutes until a tracking ("T") or a communication ("C") task starts. For two reasons, we left this sector out of our coding scheme of areas of interest. First, the sector is related to two different tasks and, hence, gazes directed at this area are difficult to interpret. Second, prompts regarding both tasks—tracking (see above) and communications (see below)—were also evident simply by looking at Sector 3 or by listening to what was communicated via the headphones. In other words, there is no strict necessity to attend to the scheduling area to perform those two tasks.

Communication: This task concerns the area in the bottom left of the displays, inside the yellow border in Figure 3, designated by the number 4. Participants wore headsets during the experiment. Through the headphones, they occasionally heard "calls" from the MATB-II. Foreign callsigns should be ignored, but if the participants' callsign was heard ("NASA504" for each participant), participants had to respond by changing the frequency of a specific radio, as they were told via this call.

Incoming calls can also be anticipated and processed via attending to the "Scheduling" zone. The green bars on the left above the "C" (=Communication) signal indicate to the participant that calls are possible to come in during these critical periods, while phases indicated by the thin orange line only mean that they cannot receive any calls during these periods and can safely ignore audio.

(Dynamic) Resource Management: In this task, which concerns the area inside the violet boundary in Figure 3, designated by the number 5, participants have to keep the filling of the Tanks "A" and "B" above and below defined thresholds (between 2000 and 3000). The tanks are slowly emptying, and participants have to use the pumps (the small areas between the tanks numbered 1 to 8) to refill or empty Tanks "A" and "B". Tanks "E" and "F" have an unlimited amount of fuel, so it is not possible to run out of fuel altogether during the test. Pumps can be activated and deactivated by a single left mouse click. Pumps can also fail (e.g., Pump 1 in Figure 3), which is signaled by the pump turning red. Participants cannot use currently failing pumps, but the failures end after some time.

Participants were randomly assigned to either a low- or high-workload group during the MATB-II block. Within the same total period of time, those participants in the low-workload group had only 3/4 the number of subtasks compared to those in the high-workload group. The total duration of the MATB-II was 5 min in both groups. To make sure that participants understood the tasks of the MATB-II, each group had a 1.5 min trial or practice run before data collection.

The MATB-II ended with the NASA Task-Load Index (NASA-TLX, Figure 4), which asked the participants about their subjectively felt workload during the task. Each subscore came with a verbal clarification. Mental Demand—"Wie mental/geistig anstrengend waren die Aufgaben?" ("How mentally demanding was the task?"). Physical Demand—"Wie physisch/körperlich anstrengend waren die Aufgaben?" ("How physically demanding was the task?"). Temporal Demand—"Wie stressing waren die Aufgaben?" ("How hurried or rushed was the pace of the task?"). Performance—"Wie gut schätzt du deine Leistungen in den Aufgaben ein?" ("How successful were you in accomplishing what you were asked to do?"). Effort—"Wie sehr musstest du dich anstrengen, um diese Leistung zu erbringen?" ("How hard did you have to work to accomplish your level of performance?"). Frustration—"Wie unsicher, entmutigt, irritiert, gestresst oder genervt hast du dich während der Aufgaben gefühlt?" ("How insecure, discouraged, irritated, stressed, or annoyed were you during the task?").

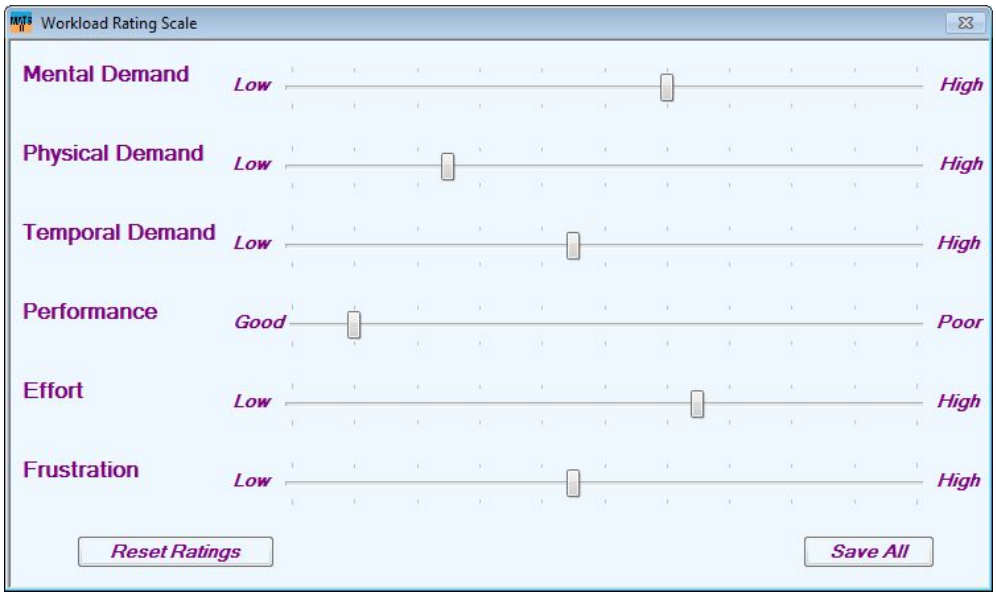


**Figure 4.** NASA Task-Load Index (NASA-TLX).

## 3. Results

### 3.1. Analyses of Bottom-Up and Top-Down Capture of Visuospatial Attention in the Visual-Search Task

For the analysis of top-down and bottom-up attention capture, only trials with correct responses were analyzed. Incorrect responses were excluded. Overall, at least 85% of the responses were correct for each participant, while the mean of correct responses was 95% (see Supplementary Table S4).

Two types of scores were calculated: a bottom-up score, which was calculated by subtracting the mean reaction time of all correct responses in the distractor-absent/target-singleton trials, in which no singleton-distractor was presented, from the mean reaction time of all correct responses in the target-dissimilar singleton-distractor trials; and a top-down score, which was calculated by subtracting the mean reaction time of all correct responses in the target-dissimilar singleton-distractor trials from the mean reaction time of all correct responses in the target-similar singleton-distractor trials. A dependent *t*-test of all 49 participants between the bottom-up scores ($M$ = 13 ms, $SD$ = 41 ms) and the top-down scores ($M$ = 56 ms, $SD$ = 83 ms) showed a significantly lower bottom-up than top-down score, $t(48) = -2.86$, $p = 0.006$, before the MATB-II task. Similar results were shown after the MATB-II task, with a slight numerical decrease in the bottom-up score ($M$ = 7 ms, $SD$ = 35 ms) as well as in the top-down score ($M$ = 37 ms, $SD$ = 56 ms), $t(48) = -2.59$, $p = 0.013$. However, no significant differences were found between the bottom-up scores before and after the MATB-II, $t(48) = 0.81$, $p = 0.421$, or between the top-down scores before and after the MATB-II, $t(48) = 1.87$, $p = 0.077$. These results are in line with those achieved in past research [8].

Next, we calculated linear regressions showing the correlations between bottom-up scores at measurement Time Points 1 and 2 and between top-down scores before and after the MATB-II. Bottom-up scores achieved before the MATB-II correlated significantly with bottom-up scores achieved after the MATB-II, $F(1, 47) = 4.245$, $p = 0.048$, $R^2 = 0.065$, adj. $R^2 = 0.045$. The same was true for top-down scores achieved before and after the MATB-II, $F(1, 47) = 22.75$, $p < 0.001$, $R^2 = 0.326$, adj. $R^2 = 0.312$. Spearman's Rho correlations were, thus, in the same ballpark as in previous studies, with bottom-up scores before and after the MATB-II, $r_s = 0.28$, $p = 0.046$; and with top-down scores before and after the MATB-II, $r_s = 0.57$, $p < 0.001$. See Figure 5.

When correlating the bottom-up scores to the top-down scores before the MATB-II, a significant but negative correlation was found, $F(1, 47) = 7.080$, $p = 0.011$, $R^2 = 0.131$, adj. $R^2 = 0.112$. Again, the same is true when correlating bottom-up scores and top-down scores after the MATB-II, $F(1, 47) = 22.130$, $p < 0.001$, $R^2 = 0.320$, adj. $R^2 = 0.306$. Spearman's Rho correlations between the bottom-up scores and the top-down scores before and after the MATB-II were as followed: $r_s = -0.45$, $p = 0.001$ before the MATB-II; and $r_s = -0.54$, $p < 0.001$ after the MATB-II. See Figure 6.

Lastly, we compared the number of times participants' gaze was distracted by the singleton distractor with the number of times participants were distracted by the nonsingletons (the gray discs). This was done to check if the distractors indeed captured attention or if longer response times (RTs) in trials with a singleton distractor reflected a nonspatial filtering cost [11]. Looking at all distractor-present trials, those with a nonmatching (target-dissimilar) singleton distractor as well as those with a matching (target-similar) singleton distractor present, we observed that participants' eyes fixated the singleton distractors more often ($M$ = 22.06, $SD$ = 13.77) than the gray nonsingletons ($M$ = 9.16, $SD$ = 8.16), $t(215) = 11.81$, $p < 0.001$. This was true despite the fact that there was only one singleton distractor in each such display, but five nonsingletons, increasing the likelihood of chance fixations on one of the gray nonsingletons discs compared to the singleton distractors.

Specifically, in nonmatching conditions, with a target-dissimilar singleton distractor, on average, participants fixated the target-dissimilar singleton distractor about 15 times ($M$ = 14.65, $SD$ = 6.96) and the nonsingletons about nine times ($M$ = 8.55, $SD$ = 8.12), $t(107) = 5.90$, $p < 0.001$. In matching conditions, with a target-similar singleton distractor present, participants fixated the target-similar singleton distractor about 30 times ($M$ = 29.46,

*SD* = 14.87) and the nonsingletons about 10 times (*M* = 9.77, *SD* = 8.15), $t(107) = 12.01$, $p < 0.001$. For a more detailed description, see Table 1.

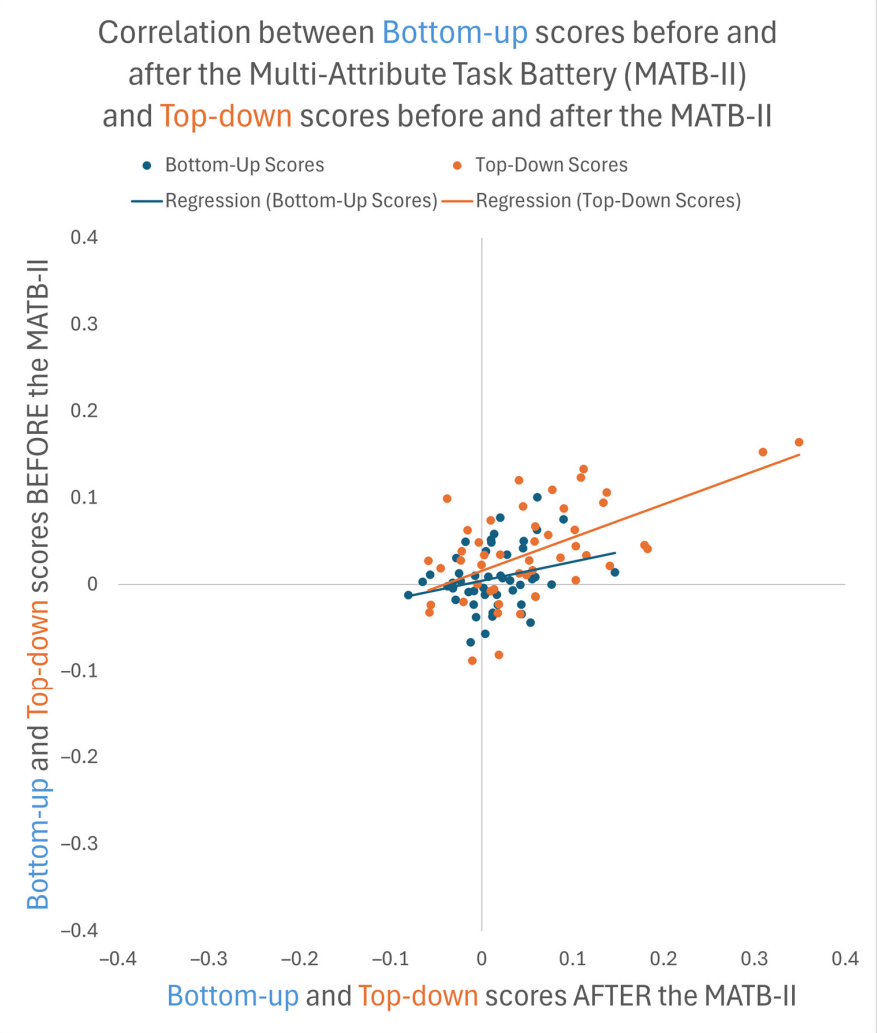


**Figure 5.** Linear regressions of bottom-up scores after the Multi-Attribute Task Battery (MATB-II) on bottom-scores before the MATB-II, as well as between top-down scores after the MATB-II on top-down scores before the MATB-II. Blue dots and orange dots correspond to individual bottom-up and top-down scores in the visual search task, respectively.

**Table 1.** Mean fixations of singleton and nonsingleton distractors divided into target color and distractor trial.

| Target-Color | Distractor Trial | Mean Target Fixation | SD Target Fixation | Mean Nonsingleton Fixation | SD Nonsingleton Fixation | Degrees of Freedom | T-Value | *p*-Value |
|---|---|---|---|---|---|---|---|---|
| Green-1 (G1) | Dissimilar<br>Similar | 18.07<br>19.75 | 8.41<br>10.85 | 12.03<br>13.25 | 10.30<br>10.35 | 27<br>27 | 1.87<br>2.25 | 0.038<br>0.028 |
| Green-2 (G2) | Dissimilar<br>Similar | 16.17<br>20.67 | 5.85<br>7.18 | 7.26<br>10.57 | 6.49<br>8.02 | 29<br>29 | 5.50<br>5.06 | <0.001<br><0.001 |
| Red-1 (R1) | Dissimilar<br>Similar | 11.85<br>40.08 | 5.25<br>9.97 | 7.00<br>5.73 | 6.35<br>4.98 | 25<br>25 | 2.94<br>15.41 | 0.005<br><0.001 |
| Red-2 (R2) | Dissimilar<br>Similar | 14.13<br>40.29 | 6.97<br>15.49 | 7.79<br>9.08 | 7.55<br>5.82 | 23<br>23 | 2.95<br>9.04 | 0.005<br><0.001 |

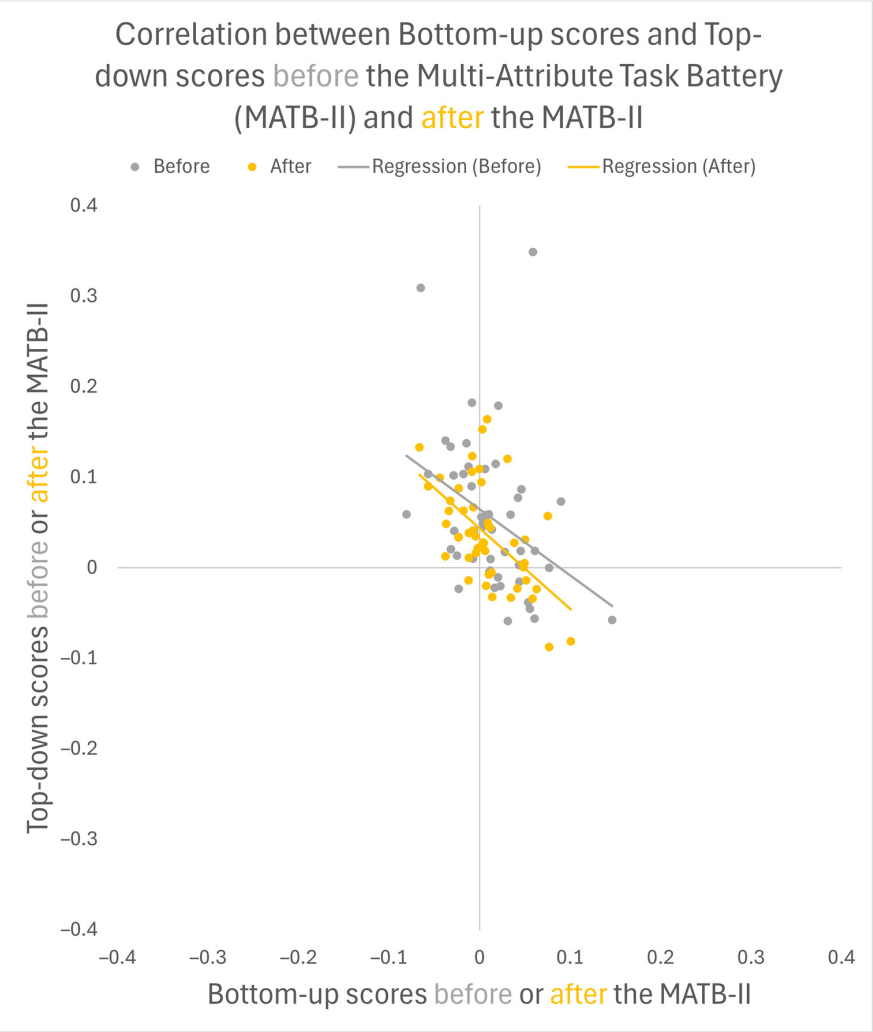


**Figure 6.** Linear regressions between bottom-up scores and top-down scores before the Multi-Attribute Task Battery (MATB-II), depicted in gray, as well as after the MATB-II, depicted in orange. Gray dots represent individual performance scores on the visual search task before the MATB-II; orange dots represent individual performance scores after the MATB-II.

Since top-down attention-capture scores were calculated as differences in RTs between matching distractor-present trials and nonmatching distractor-present trials, a correlation between the top-down scores and the fixation differences was in order. For the latter, we used the difference between the numbers of times participants looked at a matching singleton distractor versus at a nonmatching singleton distractor, once before and once after the MATB-II (see Figure 7).

Both before and after the MATB-II, we can see a clear correlation between the top-down score of the participants and the difference in the numbers of times participants fixated the matching distractors minus the nonmatching distractors (before the MATB-II: $F[(1, 47) = 27.93$, $p < 0.001$, $R^2 = 0.373$, adj. $R^2 = 0.359$, $r_s = 0.63$, $p < 0.001$; after the MATB-II: $F(1, 47) = 16.27$, $p < 0.001$, $R^2 = 0.257$, adj. $R^2 = 0.241$, $r_s = 0.55$, $p < 0.001$). No such correlations were found between the bottom-up attention-capture score and the same differences between the numbers of times participants fixated the matching singleton distractors minus the nonmatching singleton distractors, both before and after the MATB-II.

One reviewer observed that significant correlations might have been suggested by outliers only. Thus, we repeated the regression analyses without outliers. By using the interquartile range method, we identified two potential outliers—that is, participants with scores higher than the third quartile (Q3) plus 1.5-times the interquartile range (IQR; Q3 + [1.5 × IQR])

in the top-down score category, and two participants who scored higher than that in the bottom-up score category. No participants scored lower than the first quartile minus 1.5-times the interquartile range (Q1 − [1.5 × IQR]).

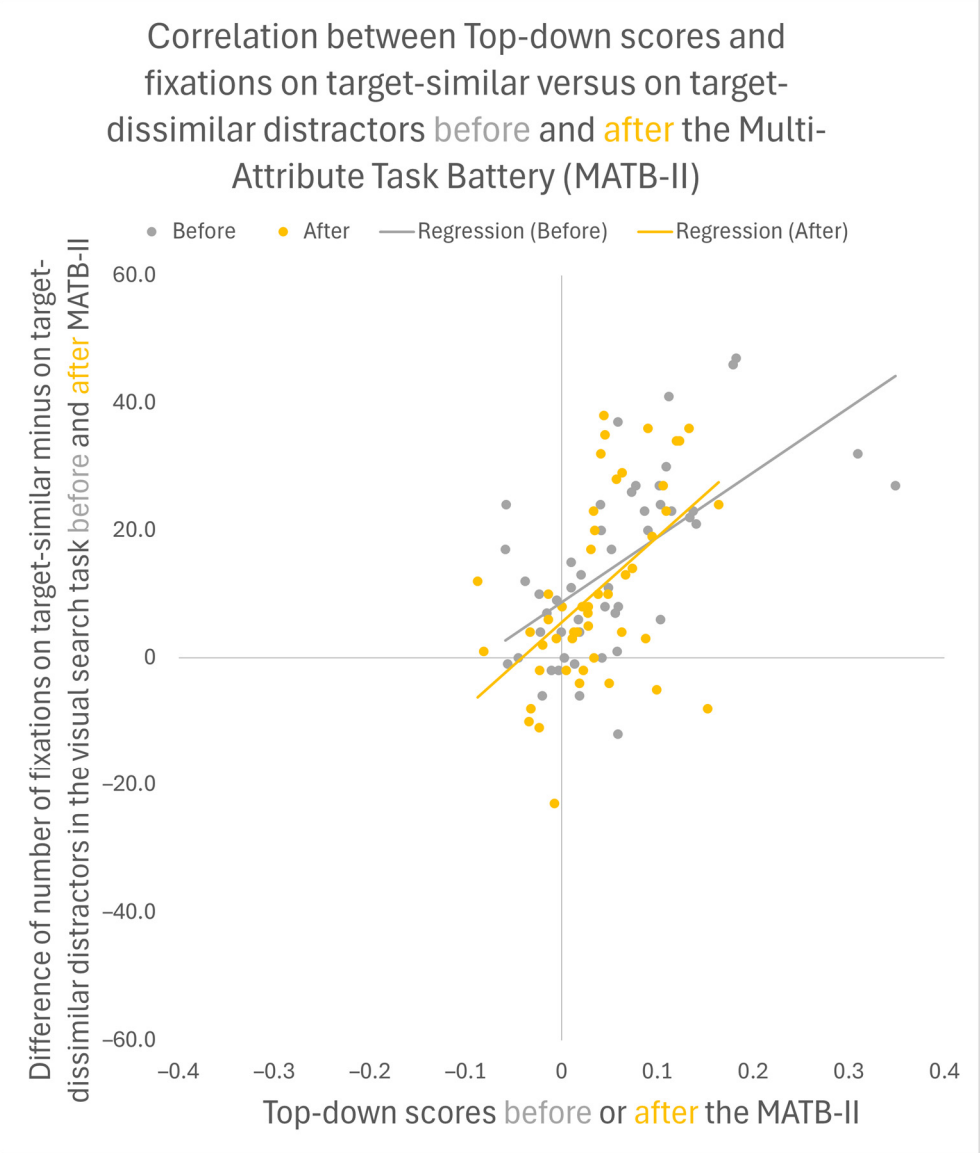


**Figure 7.** Linear regression of top-down scores on the difference between the number of fixations on target-similar minus on target-dissimilar distractors. The gray dots and the orange dots depict individual data from the visual search task before and after the Multi-Attribute Task Battery II (MATB-II), respectively.

After removing these outliers from the sample, the correlations remained significant, except for the correlations of the bottom-up scores before and after the MATB-II that dropped to a nonsignificant $R^2 = 0.050$, adj. $R^2 = 0.038$, $p = 0.053$. In contrast, we saw a positive correlation of the top-down scores before and after the MATB-II, $R^2 = 0.138$, adj. $R^2 = 0.118$, $p = 0.012$, a negative correlation between the bottom-up scores and the top-down scores before the MATB-II, $R^2 = -0.166$, adj. $R^2 = -0.146$, $p = 0.006$, and a negative correlation between the bottom-up scores and the top-down scores after the MATB-II, $R^2 = -0.340$, adj. $R^2 = -0.324$, $p < 0.001$. We also observed a positive correlation between the top-down score and the fixations on target-similar minus target-dissimilar distractors before the MATB-II, $R^2 = 0.486$, adj. $R^2 = 0.474$, $p < 0.001$, and a positive correlation between

the top-down score and the fixations on target-similar minus target-dissimilar distractors after the MATB-II, $R^2 = 0.343$, adj. $R^2 = 0.327$, $p < 0.001$.

### 3.2. Correlations between Visuospatial Attention in the Experimental Task and Performance on the Multi-Attribute Task Battery

In a next step, we correlated the bottom-up and top-down scores to the overall MATB-II scores, looking for potential correlations between scores in the MATB-II and a top-down or bottom-up capture effect, and came to the following results. A simple regression was used to regress participants' overall score in the MATB-II on the bottom-up capture effect on the one hand, and on the top-down score on the other hand. The bottom-up capture score did not predict MATB-II performance, $r(47) = -0.196$, $p = 0.177$. The same holds true for the top-down score, $r(47) = 0.152$, $p = 0.297$.

A further in-depth analysis of all participant groups (including all participants, only high-workload participants, only low-workload participants, only pilots, and all participants without pilots) regressing MATB-II scores on attention-capture scores was conducted. No significant correlation between either bottom-up or top-down scores and any of the MATB-II subtask performance scores was found, including the system monitoring performance score (see Supplementary Tables S5 and S6).

### 3.3. Analyses of Pupillary Responses in the Multi-Attribute Task Battery (MATB-II)

The analysis of the pupillary responses was conducted, with the two variables workload (high, low) and performance (high, low). As explained, we used an algorithm to automatically extract a workload measure. Figure 8 shows an example of how this works. Here, one can see how the pupil size of a participant changed during the MATB-II's subtasks, as well as the calculated change in cognitive load, and the calculated light changes that were modelled via the Pupillary-Light-Response (PLR) model [25,26]. This model predicts the pupillary light reflex behavior to brightness via an individually trained empirical model. The model uses brightness measures from the eye tracker's world video data. The algorithm uses the modelled PLR to subtract it from the raw data and arrive at a cognitive load measure.

The top left of Figure 8 shows the changes in pupil diameter over the course of the Multi-Attribute Test Battery (MATB-II; 5 min). Here, the black line corresponds to the pupil diameter in pixels. The colored vertical lines indicate the start and stop of a subtask in the MATB-II. Here, the dotted green line corresponds to a change in the pump's function in the resource-management task ("pump error" = pump cannot be used; "pump repaired" = pump can be used again). A dotted yellow line represents a change in the tracking task ("manual" = participant must control the crosshair; "auto" = the crosshair moves by itself). The dash-dotted blue line indicates the start of a communication task. The blue line corresponds to a participant's response to a communication task. The dash-dotted red line represents the start of a system monitoring task. The red line indicates that the participant responds to a system monitoring task.

The top right of Figure 8 shows the Pupillary-Light-Response (PLR) that was calculated on the basis of the light changes, presented at the bottom right. These PLRs were subtracted from the pupil diameter changes to derive the cognitive or workload changes. The bottom left of Figure 8 shows the cognitive or workload changes during the MATB-II.

Using the raw average pupil diameter sizes in pixels during the MATB-II, no significant differences between the 24 participants in the high workload group ($M = 34.08$, $SD = 6.59$) and the 25 participants in the low workload group ($M = 36.33$, $SD = 7.14$) were found, $t(47) = 1.14$, $p = 0.259$. The same is true for the cognitive load indices derived from the raw data: the average cognitive load of the high workload group ($M = 3.16$, $SD = 1.13$) did not differ significantly from the average cognitive load of the low workload group ($M = 3.46$, $SD = 1.21$), $t(47) = 0.89$, $p = 0.380$. There were also no significant differences in average raw pupil size diameter during the MATB-II between the 24 low performers ($M = 36.22$, $SD = 5.79$) and the 25 high performers ($M = 34.27$, $SD = 7.81$), $t(47) = 0.99$, p = 0.328, and

between average cognitive load measures derived from raw pupil sizes between the 24 low performers ($M$ = 3.39, $SD$ = 1.10) and the 25 high performers ($M$ = 3.25, $SD$ = 1.26), $t(47) = 0.42$, $p = 0.678$. This latter finding is particularly striking, as clear performance differences and accompanying self-assessments of felt cognitive load or workload were found between the two groups of high versus low performers (see Supplementary Table S1). This means that the average of measured pupil sizes across a task battery such as the MATB-II per se is not very revealing. Among the reasons are possible differences between different tasks and between high versus low performers concerning the rates of saccades and, thus, their contributions to measured pupil sizes. In general, saccades tend to diminish pupil size responses to task characteristics [33]. In addition, pupil size differs for different stages of task-related processing: responses before and after task-related decisions differ [34], such that even a condition-specific average pupil size measure (let alone a measure across conditions) that averages across different processing stages of task performance would not reflect cognitive load or workload, task demands, or overall performance in a complex cognitive test battery. To note, participants have to take a decision prior to each of the overt responses (e.g., a decision to start tracking the moving cursor, or a decision to press a button that turned from green to gray).

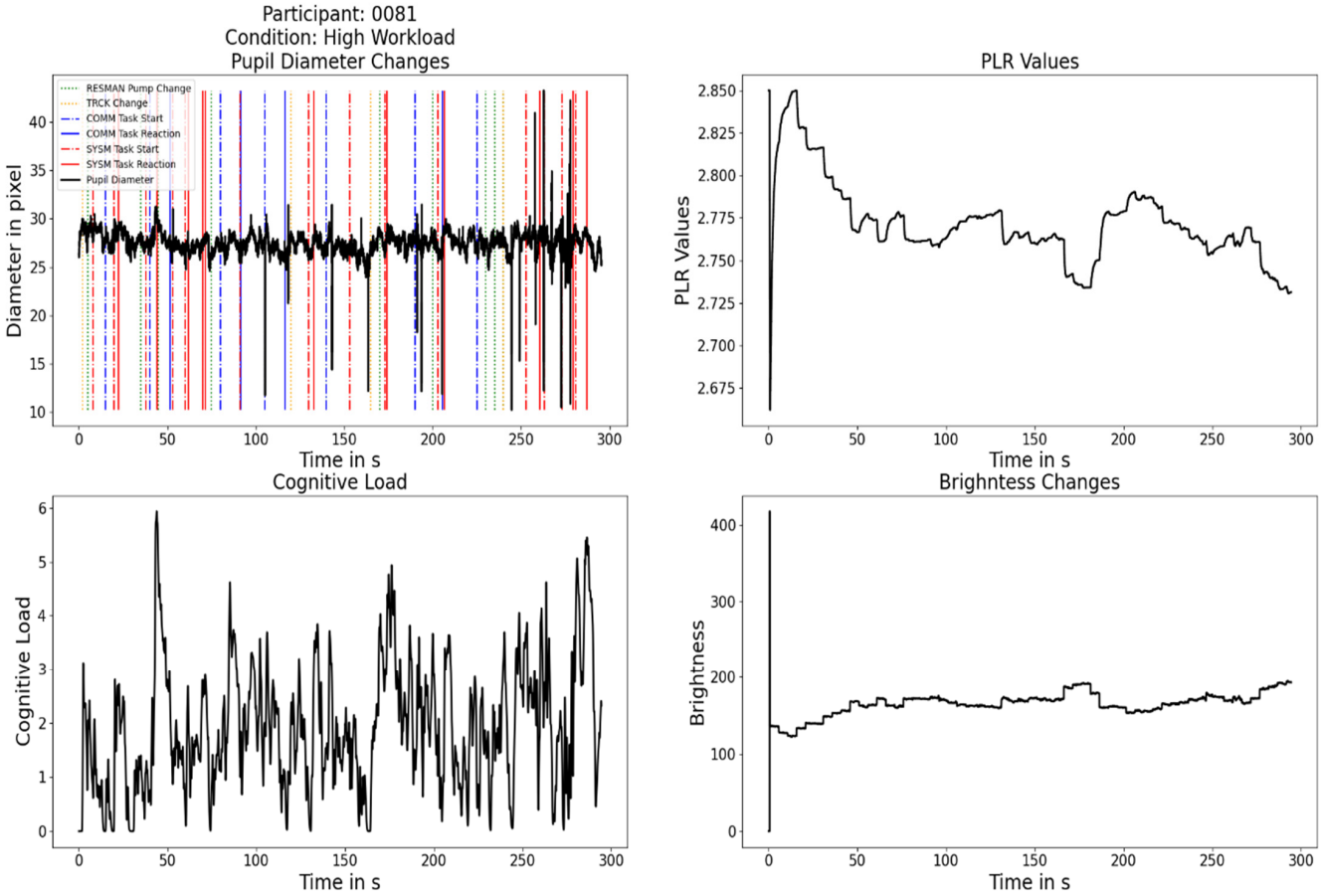


**Figure 8.** Changes in pupil size and cognitive load of a single participant in the high workload condition.

### *3.4. Analyses of Pupillary Responses in the System Monitoring Task of the Multi-Attribute Task Battery (MATB-II)*

We also analyzed performance in the system monitoring task of the MATB-II more closely because this task has potentially the tightest connection to our experimental measures of top-down versus bottom-up capture. In a first step and in direct continuation of the discussed problems regarding the usage of an average pupil size measure, we analyzed pupillary responses in this task as a function of two stages of the task. For each participant,

a "task onset" was defined by the first fixation within the task-specific area of interest following the change of the color of one of the lights/buttons, and a "task response" was defined as the moment at which each participant pressed the light/button (following a color change). The corresponding pupil sizes at these two points in time during the system monitoring task were then evaluated for whether a significant change in diameter could be found between task onset and task response. This was the case: pupil sizes at the time of task onset ($M$ = 34.29, $SD$ = 7.48) were significantly smaller than pupil sizes at the time of task response ($M$ = 37.45, $SD$ = 8.19), $t(603) = 19.02$, $p < 0.001$, analogous to past findings [34] showing that pupil size changes reveal the time point of the decision (which has to be taken before the overt manual response is given).

## 4. Discussion

In the current study, we investigated potential links between experimental measures of top-down and bottom-up capture of visuospatial attention and performance in one cognitive task battery: the MATB-II, a task battery based on operations in a simplified flight console [8]. We did so to understand if one or the other type of directing visuospatial attention—goal-directed, top-down-dependent capture of attention based on matches between visual input and searched for target features [11,35,36]; or bottom-up capture of attention due to the salience or local feature contrast between a visual singleton object and surrounding nonsingletons [13,37–39]—explains performance in the MATB-II in general or in its system monitoring subtask in particular. As performance in the MATB-II requires frequent shifting of attention between different areas on the simulated console (i.e., the monitor), and as especially the system monitoring task could be based on search for changes of lights/buttons to particular colors (i.e., from green to gray and from gray to red), we hypothesized that visuospatial attention could contribute to performance in the MATB-II or its system monitoring subtask. In addition, prior research has shown that both top-down and bottom-up scores of attention capture correlated across time and were negatively correlated with one another, meaning that the experimental task measures of visuospatial attention were not only internally valid, but also relatively stable [8,9].

### *4.1. Correlation Analyses between Visual-Search Task Measure of Visuospatial Attention and Multi-Attribute Task Battery Scores and between Different Measures of Visuospatial Attention within the Visual Search Task*

In the present study, we did not find any of the expected correlations between attention capture and MATB-II task performance. This was the case for the correlations between all (bottom-up and top-down) attention-capture scores and overall MATB-II performance, as well as for the performance scores derived from the different subtasks of the MATB-II. In addition, we replicated the positive correlations between top-down capture scores at measurement Time Points 1 and 2 (here, before and after the MATB-II) and between bottom-up capture scores at measurement Time Points 1 and 2, as well as the negative correlations between top-down and bottom-up capture scores at Time Point 1 and at Time Point 2 [8,9]. Arguably, these correlations provided the upper limit for what could be expected in terms of maximal correlations between attention capture scores and MATB-II task performance because, theoretically, it was to be expected that the correlations between the scores from one and the same task—here, the visual search task—were higher than those between the scores from two different tasks—here, the visual search task and the MATB-II—as the different tasks had less shared sources of performance variance in common [40]. This means underlying psychological functions were more different regarding performance in the visual search task versus the MATB-II than within the visual search task. For example, the task-shifting requirements [41,42] were likely higher in the MATB-II task, whereas suppression of predictable color distractors played probably a larger role in the visual search task [30,31,43]. In the current study, by blocking the different distractor conditions in the visual search task (e.g., by presenting first all trials without a singleton distractor, then all trials with a nonmatching/target-dissimilar distractor, and then all trials with a matching/target-similar distractors), we reduced the residual shift costs that might have

theoretically resulted from trying to ignore different specific color singletons from one trial to the next [44,45]. In contrast, task shifts are common in the MATB-II [46]. Likewise, by using different colors to indicate different tasks (e.g., changes from green to gray and from gray to red in the system monitoring tasks) of the MATB-II, we made it very difficult if not impossible for participants to proactively suppress particular irrelevant colors in the test battery [47]. In contrast, using the same singleton-distractor colors for blocks of trials, proactive suppression of distractor interference was probably a factor in the visual search task [31].

Summarizing, it seems that the chances for finding any significant correlations between the experimental visual search task's attention capture scores and the MATB-II performance scores were limited from the start by the less than perfect reliability or temporal consistency (i.e., the correlations $< 1.00$) of the attention-capture scores in the first place, especially the bottom-up capture score. In addition, there might have been good theoretical reasons why the correlations between attention capture scores and MATB-II performance were low or nonexistent. For example, in the current study, the correlations of the top-down scores were numerically not as low across measurement time points as that of the bottom-up capture scores. Thus, theoretically, there was more space for a correlation of the top-down capture score with the MATB-II performance, especially in a subtask such as system monitoring that participants could have solved by searching for specific colors. However, potential correlations between top-down capture scores and MATB-II task performance that were currently not found could have suffered for theoretical reasons alone. For example, past research has shown that participants have increasing difficulty to proactively search for several relevant colors at the same time [48–52]. However, this was what was required in the MATB-II task. For example, a change from a green to a gray color and a change from a gray to a red color were both task-relevant in the system monitoring task of the MATB-II. In fact, with its fixed positions on the monitor, participants in the MATB-II task could have even used a location-based monitoring (or search) strategy for their task-shifting and subtask performance [53–55]. For example, it is known that participants can exploit their knowledge of likely locations of objects in a scene for their shifts of attention and their eye movements [56–58]. In the present study, considering that specific colors (e.g., the color red) had different meanings depending on where they were located in the display (e.g., they indicated that a button had to be pressed in Areas 1 and 2, see Figure 3, but that a pump failed and can currently not be used to replenish tanks in Area 5, see Figure 3), a location-based strategy or a strategy that looks for conjunctions of specific colors and locations is not unlikely to account for performance in the MATB-II.

### *4.2. Further Findings of Interest*

In addition to these most important findings, we observed surprising effects of cognitive load (or workload) on performance in the MATB-II (see Supplementary Table S2). To increase variance in the MATB-II performance, we used two workload conditions differing in the number of tasks and task shifts per unit of time. Contrary to what we would have expected, however, in the present study, participants' performance was higher in the system monitoring and tracking subtasks under the high- than under the low-workload conditions. Typically, performance in cognitive tasks such as the MATB-II declines with a higher workload [46,59,60]. There are several possible reasons for the presently found deviation from this expected pattern. First, participants in the low-workload condition might not have performed close to capacity, meaning that there could have been spare capacity to prioritize the two subtasks of system monitoring and of tracking, for which we found performance improvements relative to the low-workload conditions. Second, because we used a between-participants design, it is possible that generally better performing participants were placed in the high-workload than in the low-workload group. Third, somewhat related to the first point, general physiological activation in the less demanding conditions might have been too low for optimal performance. It is assumed that emotions experienced in "boring" tasks, imposing too little demands, and the

resulting achievement motivation could be too low for optimal cognitive performance [61]. Partly in line with this proposal, early findings suggested, for instance, an inversely u-shaped function relating physiological arousal to task performance [62]. Fourth, a very interesting possibility has recently been suggested when dual-task performance benefits over single-task performance were observed [63,64]. Researchers [63] believe that the necessity to suppress a prepotent response that would naturally occur under some single-task conditions (e.g., not being allowed to look at a target when a spatially compatible manual response to the target is required) could create a single-task cost that is absent when participants are allowed to perform both responses (or "tasks"). According to this reasoning, subtasks in the MATB-II could partly better be integrated with one another in the high-workload condition than in the low-workload condition. For example, alluding to the possibility of a location-based search strategy to look for changes of the stimuli that we discussed above, a higher frequency of the system monitoring task in the high-load conditions could have led to more fixations in this area of the screen (Areas 1 and 2, see Figure 3), and performance on the visually controlled task in the spatially adjacent area (Area 3, see Figure 3, here, the tracking task) could have benefitted from this general looking-direction effect. This type of coupled benefit between performance on these two tasks would be fully in line with our observation of a better performance in these two tasks under high-load conditions (see Supplementary Table S2). It is also in line with a post hoc comparison of the overall higher fixation durations on regions of interest of the tracking tasks under low- than under high-load conditions, suggesting that a higher rate of switching between tasks allowed the participants to be more aware of changes that needed a participant's response. Of the 24 participants in the high-workload condition, only a mean of 36% ($SD$ = 9.64) of the time spent looking at the MATB-II tasks in total was directed at the area of interest of the tracking task, while the 25 participants of the low-workload condition spent a significantly larger mean amount of 42% ($SD$ = 8.75) fixating on the tracking task, $t(49) = 2.09$, $p = 0.047$. Other subtasks of the MATB-II under high-workload conditions might have neither benefited, nor suffered from this dual-task benefit for system monitoring and tracking because these other tasks relied on auditory input (i.e., the communication task) and, thus, would not benefit from visuospatial attention being directed to an adjacent area, or were less dependent on directing visuospatial attention to the particular region of the monitor for other reasons such as being relatively insensitive to the exact time at which the task was handled. The latter would have been the case for the resource management task, for which we did not even analyze reaction times, and which was also least sensitive of all subtasks to the performance difference between high- and low-performers in the MATB-II (see Supplementary Table S7).

### *4.3. Pupillary Cognitive Load or Workload Responses*

A related point of interest concerns the insensitivity of our pupillary cognitive load index to the manipulations of workload, but also to the factual task performance—that is, to the median split of our participants into high versus low performers on the MATB-II. To note, the cognitive load index is a computationally modelled load-elicited pupillary response that is supposed to be free of the luminance-elicited pupillary size change. While we could relatively easily explain the lacking impact of our workload manipulation on the cognitive load index of the pupillary response through the lacking predicted impact of our workload manipulation on performance in the MATB-II, this is not the case for the absence of a difference in the cognitive load index of the pupillary response between high and low performers. The latter groups clearly differed from one another in terms of their performance in the MATB-II in all but one subtask (i.e., the resource management task). Yet, these groups did not differ with respect to their cognitive load index derived from pupillary responses. In addition, high versus low performers also differed regarding their self-assessed workload (see Supplementary Table S8). These significant differences imply that there would have been theoretical reasons to expect a difference in the cognitive load index of the pupillary response. At least three possibilities come to mind explaining

the discrepancy between the two measures (objective performance on the MATB-II versus pupil-based load index). First, it is possible that averaging the cognitive load index of the pupillary response across different stages of the subtasks of the MATB-II simply washed out any load differences between high and low performers due to averaging across times of very different sensitivity of the pupils to the load differences. For instance, recent research suggested that indices of pupil sizes could vary depending on the amount of saccades conducted in a task [33]. This was not controlled for in the current study. This general possibility of a watered-down effect of averaging would also be in line with the general observation of stages of different sensitivities of the pupil size for task demands, such as pre- versus postdecisional stages [34]. This possibility of averaging out of the pupillary response to varying degrees would also be supported by the following observation: we observed a difference in the pupillary response at task onset versus at task response. This difference came to light in our more detailed analysis of the system monitoring task performance. Secondly, other studies have also found that different measures of cognitive load such as task performance and pupillary responses do not always converge [60]. In fact, even the performances on different visual tasks that are meant to measure the same aptitude do not necessarily converge [40]. Thirdly, we believe that there is also space for improvement of the estimation of the light-elicited response in spatially articulated displays such as that of the currently used MATB-II. For example, it is known that even subjectively perceived lightness can prompt a light-elicited pupillary response [65,66]. There is arguably room for such illusory lightness in articulated monitor displays [67] such as the ones used in the MATB-II that is currently not ruled out by the automatic measurement and subtraction of the objective light-elicited pupillary response [25,26], so that it is possible that the corresponding artifacts in the pupil size measures could have watered down the true influence of cognitive load on pupil size in the present study's MATB-II, too.

### *4.4. Limitations*

Our discussion already revealed a number of limitations. We pointed out that future studies should take saccade rates into account and that they need to carefully discriminate not only between tasks, but also between stages of task-specific processing to make use of pupil-size measures. In addition, we argued that pupil size measures of workload may also benefit from taking into account more subtle visual brightness effects than are currently measured with the video camera-based brightness measurement. However, we think that controlling for additional influences on pupil size that are relatively independent of workload and brightness (e.g., emotions) is not necessary, unless one has good reasons to assume that such independent effects are confounded with effects of workload.

In addition to these points, more generally, we based our sample size estimates on substantial effect sizes. Certainly, weaker effects were, therefore, impossible for us to detect. Regarding the participants that we tested, we failed to collect more data from experts, in our case, pilots. Instead, our sample consisted of mostly students. This is maybe not ideal in two respects. On the one hand, more pilots could have meant that task-relevant performance variance would have increased, allowing all variance-based measures a better chance to be detected. On the other hand, students might also have been relatively good performers, meaning that especially weaker performers would have been missing, again restricting the overall performance variance in an unnecessary way.

At a theoretical level, one could argue that too little is known yet about how separate cognitive functions such as bottom-up or top-down visuospatial attention play out in more complex applied or real-world tasks. As a consequence, testing for the role in a more applied setting might have been overly optimistic in the first place. This is true, but we would also want to point out that research such as the present study would help to inform these applied or real-world task models by demonstrating if a particular cognitive mechanism or function would have to be taken into account to explain cognitive task performance, yes or no. At a conceptual level, one could also argue that the major purpose of a task battery (e.g., the MATB-II) is its usage in diagnostics. What matters most is if a

task or test (battery) could tell people with and without high aptitude from one another. In contrast, understanding the exact working of such tasks or tests would not be necessary for this purpose. Here, we would like to argue, however, that the internal validity of diagnostic tasks or tests is important, even for the practical purposes mentioned above. Knowing what exactly accounts for task or test performance could help to increase task or test sensitivity for the aptitude in question even further. For example, knowing what accounts for task performance would allow one to construct trials or items suited to measure the major performance contributors.

## 5. Conclusions

In the current study, we found no evidence that measures of top-down or bottom-up capture of visuospatial attention had any bearing on performance in a more applied cognitive task battery. Just as being good at cycling is not sufficient to perform well during a triathlon (which requires one to also run and swim well), visuospatial attention could simply not be decisive for overall performance when operating a flight console. This finding casts doubt on the generalizability of experimental task performance to more applied and real-world tasks. This finding also emphasizes the doubtful ecological validity of many experimental tasks (although they are doubtlessly of high internal validity) [68]. We could also not find any links between MATB-II task performance and cognitive load indices derived from pupillary size. Moreover, we observed surprising effects of workload (or cognitive load) manipulation on MATB-II performance itself. Maybe it is not too surprising that pupillary responses did not react to the task load manipulation because the latter created a paradoxical effect. However, we want to emphasize that there were also no significant correlations between pupillary responses and individual MATB-II performance. These findings imply that the MATB-II itself poses a number of questions about its underlying rationale. These findings also revealed that pupillary responses are not necessarily an ideal tool to tell participants of varying aptitude apart. This conclusion at least holds for the relatively homogenous sample of mostly student participants that we used in the current study. Nevertheless, these types of studies, where concepts with a strong foundation, which in our case would be the bottom-up and top-down search task paradigm, and real-world use-cases are compared toe to toe, are incredibly beneficial for our understanding of limitations of lab studies, as well as possibly finding issues in validity and reliability of a real-world use-case testing apparatus.

**Supplementary Materials:** The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app14083158/s1, Table S1: t T-Test of all pupil diameter and cognitive load changes during the Multi-Attribute Task Battery (MATB-II) between low- and high-workload groups, as well as between low- and high-performer groups, the latter also split for within the low-workload group and the high-workload group.; Table S2: T-Test of all Multi-Attribute Task Battery (MATB-II) subtasks z-scores, between participants in the low-workload group and the high-workload group, as well as between participants in the low-performance group and participants in the high-performance group.; Table S3: T-Test of all Multi-Attribute Task Battery (MATB-II) subtasks z-scores, between low- and high-performing participants within either the low-workload group or the high-workload group.; Table S4: List of all participants' search-task scores.; Table S5: Pearson Correlation coefficients between Multi-Attribute Task Battery (MATB-II) sub-task performance scores and bottom-up attention-capture scores divided by groups (all participants, all participants - pilots excluded, low-workload group, high-workload group, pilots only); Table S6: Pearson Correlation coefficients between Multi-Attribute Task Battery (MATB-II) sub-task performances and top-down attention-capture scores divided by groups (all participants, all participants - pilots excluded, low-workload group, high-workload group, pilots only); Table S7: X-Y coordinate tracking. Percentages of which area eyes were directed at.; Table S8: T-Test of all NASA Task Load Index (TLX) scores, between participants in the low-workload group and the high-workload group, as well as between participants in the low-performance group and participants in the high-performance group.

**Author Contributions:** Conceptualization, B.G., M.S. and U.A.; Methodology, D.G., M.S. and U.A.; Software, D.G. and B.G.; Validation, D.G. and B.G.; Formal analysis, D.G.; Resources, U.A.; Data curation, D.G.;

Writing—original draft, D.G. and U.A.; Writing—review & editing, D.G., B.G., M.S. and U.A.; Visualization, D.G.; Supervision, B.G., M.S. and U.A.; Project administration, B.G. and U.A.; Funding acquisition, U.A. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by Austrian Research Promotion Agency grant number [880102]. Open Access Funding by the University of Vienna.

**Institutional Review Board Statement:** The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of University of Vienna (protocol code 00644 at the 13. of April 2021). for studies involving humans.

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Data Availability Statement:** The data presented in this study are openly available in FigShare at https://doi.org/10.6084/m9.figshare.25540156.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Kong, Y.; Posada-Quintero, H.F.; Gever, D.; Bonacci, L.; Chon, K.H.; Bolkhovsky, J. Multi-Attribute Task Battery configuration to effectively assess pilot performance deterioration during prolonged wakefulness. *Inform. Med. Unlocked* **2022**, *28*, 100822. [CrossRef]
2. Wang, P.; Fang, W.; Guo, B. A colored petri nets based workload evaluation model and its validation through Multi-Attribute Task Battery-II. *Appl. Ergon.* **2017**, *60*, 260–274. [CrossRef] [PubMed]
3. Fukuda, K.; Vogel, E.K.; Awh, E. Quantity, not quality: The relationship between fluid intelligence and working memory capacity. *Psychon. Bull. Rev.* **2010**, *17*, 673–679. [CrossRef] [PubMed]
4. Robertson, I.H.; Ward, T.; Ridgeway, V.; Nimmo-Smith, I. The structure of normal human attention: The Test of Everyday Attention. *J. Int. Neuropsychol. Soc.* **1996**, *2*, 525–534. [CrossRef]
5. Roque, N.A.; Wright, T.J.; Boot, W.R. Do different attention capture paradigms measure different types of capture? *Atten. Percept. Psychophys.* **2016**, *78*, 2014–2030. [CrossRef] [PubMed]
6. Conway, A.R.A.; Kane, M.J.; Bunting, M.F.; Hambrick, D.Z.; Wilhelm, O.; Engle, R.W. Working memory span tasks: A methodological review and user's guide. *Psychon. Bull. Rev.* **2005**, *12*, 769–786. [CrossRef]
7. Wiegand, I.; Töllner, T.; Habekost, T.; Dyrholm, M.; Müller, H.J.; Finke, K. Distinct Neural Markers of TVA-Based Visual Processing Speed and Short-Term Storage Capacity Parameters. *Cereb. Cortex* **2014**, *24*, 1967–1978. [CrossRef] [PubMed]
8. Weichselbaum, H.; Huber-Huber, C.; Ansorge, U. Attention capture is temporally stable: Evidence from mixed-model correlations. *Cognition* **2018**, *180*, 206–224. [CrossRef]
9. Weichselbaum, H.; Ansorge, U. Bottom-up attention capture with distractor and target singletons defined in the same (color) dimension is not a matter of feature uncertainty. *Atten. Percept. Psychophys.* **2018**, *80*, 1350–1361. [CrossRef]
10. Santiago-Espada, Y.; Myer, R.; Latorella, K.; Comstock, J.R. The Multi-Attribute Task Battery II (MATB-II) Software for Human Performance and Workload Research: A User's Guide. 2011. Available online: https://www.semanticscholar.org/paper/The-Multi-Attribute-Task-Battery-II-(\protect\unhbox\voidb@x\hbox{MATB-II})-for-A-Santiago-Espada-Myer/03048e4a70abc42693148a7b4e24d2a18ab75347 (accessed on 29 August 2023).
11. Folk, C.L.; Remington, R. Selectivity in distraction by irrelevant featural singletons: Evidence for two forms of attentional capture. *J. Exp. Psychol. Hum. Percept. Perform.* **1998**, *24*, 847–858. [CrossRef]
12. Goller, F.; Ditye, T.; Ansorge, U. The contribution of color to attention capture effects during search for onset targets. *Atten. Percept. Psychophys.* **2016**, *78*, 789–807. [CrossRef] [PubMed]
13. Theeuwes, J. Cross-dimensional perceptual selectivity. *Percept. Psychophys.* **1991**, *50*, 184–193. [CrossRef] [PubMed]
14. Theeuwes, J. Perceptual selectivity for color and form. *Percept. Psychophys.* **1992**, *51*, 599–606. [CrossRef] [PubMed]
15. Theeuwes, J. Top–down and bottom–up control of visual selection. *Acta Psychol.* **2010**, *135*, 77–99. [CrossRef]
16. Appel, T.; Scharinger, C.; Gerjets, P.; Kasneci, E. Cross-subject workload classification using pupil-related measures. In Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications, Warsaw, Poland, 14–18 June 2018; pp. 1–8. [CrossRef]
17. Chen, S.; Epps, J. Automatic classification of eye activity for cognitive load measurement with emotion interference. *Comput. Methods Programs Biomed.* **2013**, *110*, 111–124. [CrossRef] [PubMed]
18. Chen, S.; Epps, J. Using Task-Induced Pupil Diameter and Blink Rate to Infer Cognitive Load. *Hum. Comput. Interact.* **2014**, *29*, 390–413. [CrossRef]
19. Iqbal, S.T.; Zheng, X.S.; Bailey, B.P. Task-evoked pupillary response to mental workload in human-computer interaction. In Proceedings of the CHI'04 Extended Abstracts on Human Factors in Computing Systems, Vienna, Austria, 24–29 April 2004; pp. 1477–1480.
20. Krejtz, K.; Duchowski, A.T.; Niedzielska, A.; Biele, C.; Krejtz, I. Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. *PLoS ONE* **2018**, *13*, e0203629. [CrossRef] [PubMed]
21. Laeng, B.; Ørbo, M.; Holmlund, T.; Miozzo, M. Pupillary Stroop effects. *Cogn. Process.* **2011**, *12*, 13–21. [CrossRef]

22. Stolte, M.; Gollan, B.; Ansorge, U. Tracking visual search demands and memory load through pupil dilation. *J. Vis.* **2020**, *20*, 21. [CrossRef]
23. Ahlstrom, U.; Friedman-Berg, F.J. Using eye movement activity as a correlate of cognitive workload. *Int. J. Ind. Ergon.* **2006**, *36*, 623–636. [CrossRef]
24. Van der Wel, P.; van Steenbergen, H. Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychon. Bull. Rev.* **2018**, *25*, 2005–2015. [CrossRef] [PubMed]
25. Gollan, B. Sensor-based Online Assessment of Human Attention. Ph.D. Thesis, Johannes Kepler University Linz, Linz, Austria, 2018. [CrossRef]
26. Gollan, B.; Ferscha, A. Modeling Pupil Dilation as Online Input for Estimation of Cognitive Load in non-laboratory Attention-Aware Systems. In Proceedings of the COGNITIVE 2016: The Eighth International Conference on Advanced Cognitive Technologies and Applications, Rome, Italy, 20–24 March 2016.
27. Bradley, M.M.; Miccoli, L.; Escrig, M.A.; Lang, P.J. The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology* **2008**, *45*, 602–607. [CrossRef] [PubMed]
28. Casuccio, A.; Cillino, G.; Pavone, C.; Spitale, E.; Cillino, S. Pharmacologic pupil dilation as a predictive test for the risk for intraoperative floppy-iris syndrome. *J. Cataract. Refract. Surg.* **2011**, *37*, 1447–1454. [CrossRef] [PubMed]
29. Clewett, D.; Gasser, C.; Davachi, L. Pupil-linked arousal signals track the temporal organization of events in memory. *Nat. Commun.* **2020**, *11*, 4007. [CrossRef]
30. Gao, Y.; Theeuwes, J. Learning to suppress a distractor is not affected by working memory load. *Psychon. Bull. Rev.* **2020**, *27*, 96–104. [CrossRef] [PubMed]
31. Stilwell, B.T.; Bahle, B.; Vecera, S.P. Feature-based statistical regularities of distractors modulate attentional capture. *J. Exp. Psychol. Hum. Percept. Perform.* **2019**, *45*, 419–433. [CrossRef] [PubMed]
32. Wang, B.; Theeuwes, J. How to inhibit a distractor location? Statistical learning versus active, top-down suppression. *Atten. Percept. Psychophys.* **2018**, *80*, 860–870. [CrossRef] [PubMed]
33. Burlingham, C.S.; Mirbagheri, S.; Heeger, D.J. A unified model of the task-evoked pupil response. *Sci. Adv.* **2022**, *8*, eabi9979. [CrossRef] [PubMed]
34. Einhäuser, W.; Koch, C.; Carter, O.L. Pupil dilation betrays the timing of decisions. *Front. Hum. Neurosci.* **2010**, *4*, 18. [CrossRef]
35. Bundesen, C. A theory of visual attention. *Psychol. Rev.* **1990**, *97*, 523–547. [CrossRef]
36. Duncan, J.; Humphreys, G.W. Visual search and stimulus similarity. *Psychol. Rev.* **1989**, *96*, 433–458. [CrossRef] [PubMed]
37. Itti, L.; Koch, C.; Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 1254–1259. [CrossRef]
38. Li, Z. A saliency map in primary visual cortex. *Trends Cogn. Sci.* **2002**, *6*, 9–16. [CrossRef] [PubMed]
39. Nothdurft, H.-C. Salience from feature contrast: Additivity across dimensions. *Vis. Res.* **2000**, *40*, 1183–1201. [CrossRef] [PubMed]
40. Cappe, C.; Clarke, A.; Mohr, C.; Herzog, M. Is there a common factor for vision? *J. Vis.* **2014**, *14*, 4. [CrossRef] [PubMed]
41. Miyake, A.; Friedman, N.P.; Emerson, M.J.; Witzki, A.H.; Howerter, A.; Wager, T.D. The Unity and Diversity of Executive Functions and Their Contributions to Complex "Frontal Lobe" Tasks: A Latent Variable Analysis. *Cogn. Psychol.* **2000**, *41*, 49–100. [CrossRef] [PubMed]
42. Monsell, S. Task switching. *Trends Cogn. Sci.* **2003**, *7*, 134–140. [CrossRef]
43. Stilwell, B.T.; Vecera, S.P. Learned distractor rejection in the face of strong target guidance. *J. Exp. Psychol. Hum. Percept. Perform.* **2020**, *46*, 926–941. [CrossRef]
44. Reeder, R.R.; Olivers, C.N.; Hanke, M.; Pollmann, S. No evidence for enhanced distractor template representation in early visual cortex. *Cortex* **2018**, *108*, 279–282. [CrossRef]
45. De Vries, I.E.J.; Savran, E.; Van Driel, J.; Olivers, C.N.L. Oscillatory Mechanisms of Preparing for Visual Distraction. *J. Cogn. Neurosci.* **2019**, *31*, 1873–1894. [CrossRef]
46. Gutzwiller, R.S.; Wickens, C.D.; Clegg, B.A. Workload overload modeling: An experiment with MATB II to inform a computational model of task management. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **2014**, *58*, 849–853. [CrossRef]
47. Kerzel, D.; Barras, C. Distractor rejection in visual search breaks down with more than a single distractor feature. *J. Exp. Psychol. Hum. Percept. Perform.* **2016**, *42*, 648–657. [CrossRef] [PubMed]
48. Büsel, C.; Pomper, U.; Ansorge, U. Capture of attention by target-similar cues during dual-color search reflects reactive control among top-down selected attentional control settings. *Psychon. Bull. Rev.* **2019**, *26*, 531–537. [CrossRef]
49. Folk, C.L.; Anderson, B.A. Target-uncertainty effects in attentional capture: Color-singleton set or multiple attentional control settings? *Psychon. Bull. Rev.* **2010**, *17*, 421–426. [CrossRef] [PubMed]
50. Grubert, A.; Eimer, M. A capacity limit for the rapid parallel selection of multiple target objects. *J. Vis.* **2018**, *18*, 1017. [CrossRef]
51. Kerzel, D.; Grubert, A. Capacity limitations in template-guided multiple color search. *Psychon. Bull. Rev.* **2022**, *29*, 901–909. [CrossRef]
52. Ort, E.; Fahrenfort, J.J.; Olivers, C.N.L. Lack of Free Choice Reveals the Cost of Having to Search for More Than One Object. *Psychol. Sci.* **2017**, *28*, 1137–1147. [CrossRef]
53. Pereira, E.J.; Castelhano, M.S. Attentional capture is contingent on scene region: Using surface guidance framework to explore attentional mechanisms during search. *Psychon. Bull. Rev.* **2019**, *26*, 1273–1281. [CrossRef]

54. Torralba, A.; Oliva, A.; Castelhano, M.S.; Henderson, J.M. Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychol. Rev.* **2006**, *113*, 766–786. [CrossRef]
55. Wolfe, J.M. Guided Search 6.0: An updated model of visual search. *Psychon. Bull. Rev.* **2021**, *28*, 1060–1092. [CrossRef]
56. Eckstein, M.P.; Drescher, B.A.; Shimozaki, S.S. Attentional Cues in Real Scenes, Saccadic Targeting, and Bayesian Priors. *Psychol. Sci.* **2006**, *17*, 973–980. [CrossRef] [PubMed]
57. Võ, M.L.-H.; Wolfe, J.M. Differential Electrophysiological Signatures of Semantic and Syntactic Scene Processing. *Psychol. Sci.* **2013**, *24*, 1816–1823. [CrossRef] [PubMed]
58. Võ, M.L.; Wolfe, J.M. The role of memory for visual search in scenes. *Ann. New York Acad. Sci.* **2015**, *1339*, 72–81. [CrossRef]
59. Bulikhov, D.; Landry, S.J. The effect of applied effort on MATB-II performance. *Theor. Issues Ergon. Sci.* **2023**, *24*, 233–240. [CrossRef]
60. Muñoz-de-Escalona, E.; Cañas, J.J.; Leva, C.; Longo, L. Task Demand Transition Peak Point Effects on Mental Workload Measures Divergence. In Proceedings of the Human Mental Workload: Models and Applications: 4th International Symposium, H-WORKLOAD 2020, Granada, Spain, 3–5 December 2020; Longo, L., Leva, M.C., Eds.; Communications in Computer and Information Science; Springer International Publishing: Cham, Switzerland, 2020; Volume 1318, pp. 207–226. [CrossRef]
61. Pekrun, R.; Frenzel, A.C.; Goetz, T.; Perry, R.P. The Control-Value Theory of Achievement Emotions. In *Emotion in Education*; Elsevier: Amsterdam, The Netherlands, 2007; pp. 13–36. [CrossRef]
62. Yerkes, R.M.; Dodson, J.D. The relation of strength of stimulus to rapidity of habit-formation. *J. Comp. Neurol. Psychol.* **1908**, *18*, 459–482. [CrossRef]
63. Huestegge, L.; Koch, I. When two actions are easier than one: How inhibitory control demands affect response processing. *Acta Psychol.* **2014**, *151*, 230–236. [CrossRef] [PubMed]
64. Kürten, J.; Raettig, T.; Gutzeit, J.; Huestegge, L. Dual-action benefits: Global (action-inherent) and local (transient) sources of action prepotency underlying inhibition failures in multiple action control. *Psychol. Res.* **2023**, *87*, 410–424. [CrossRef] [PubMed]
65. Laeng, B.; Endestad, T. Bright illusions reduce the eye's pupil. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 2162–2167. [CrossRef] [PubMed]
66. Laeng, B.; Nabil, S.; Kitaoka, A. The Eye Pupil Adjusts to Illusorily Expanding Holes. *Front. Hum. Neurosci.* **2022**, *16*, 877249. [CrossRef]
67. Bressan, P.; Actis-Grosso, R. Simultaneous Lightness Contrast on Plain and Articulated Surrounds. *Perception* **2006**, *35*, 445–452. [CrossRef]
68. Cavanagh, P.; Alvarez, G. Tracking multiple targets with multifocal attention. *Trends Cogn. Sci.* **2005**, *9*, 349–354. [CrossRef] [PubMed]