UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculteit Wetenschappen
School voor Informatietechnologie

master in de informatica

Masterthesis

Measuring cognitive load for complex assembly tasks in Virtual Reality

Devlin Vranken

Scriptie ingediend tot het behalen van de graad van master in de informatica

PROMOTOR :
Prof. dr. Gustavo Alberto ROVELO RUIZ

COPROMOTOR :
dr. Eva GEURTS

2024
2025

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

Maastricht University

Faculteit Wetenschappen
School voor Informatietechnologie

master in de informatica

Masterthesis

Measuring cognitive load for complex assembly tasks in Virtual Reality

Devlin Vranken

Scriptie ingediend tot het behalen van de graad van master in de informatica

PROMOTOR :
Prof. dr. Gustavo Alberto ROVELO RUIZ

COPROMOTOR :
dr. Eva GEURTS

Acknowledgments

I would like to express my sincere gratitude to Prof. Dr. Gustavo Rovelo Ruiz and Dr. Eva Geurts for their invaluable support throughout the development of my thesis. Their guidance and constructive feedback played a crucial role in shaping the direction of my research. From the early stages of conceptualization to the final execution of the user study, their insights and suggestions were instrumental. I am especially thankful for the time they dedicated to reviewing my work at multiple stages and helping me improve both the content and clarity of the thesis.

I would also like to thank Arno Verstraete for the technical support and advice provided whenever I encountered implementation challenges. His practical insights and readiness to assist helped me overcome several obstacles and kept the development process on track.

Their combined support has been essential to the successful completion of this work, and I am truly grateful for their involvement.

This research was made possible with support from the MAXVR-INFRA project, a scalable and flexible infrastructure that facilitates the transition to digital-physical work environments. The MAXVR-INFRA project is funded by the European Union - NextGenerationEU and the Flemish Government.

Summary

Introduction

This research investigates how the cognitive load of individuals can be accurately measured while performing complex assembly tasks in a virtual reality (VR) environment. With the rise of VR as a training tool in industrial contexts, such as simulating work situations on oil platforms or power plants, the need for reliable methods to measure mental workload is also growing. An accurate assessment of that workload is essential to evaluate the effectiveness of training and to avoid overload or underload.

Previous studies show that eye measurements, such as pupil size and blink frequency, can be strong indicators of cognitive load. Since many modern VR headsets are equipped with built-in eye tracking, this offers a unique opportunity to integrate these measurements naturally into virtual simulations. However, measuring cognitive load through eye data is not without challenges. Factors such as screen brightness, object colors, or the virtual depth of elements can significantly influence pupil behavior. This creates a risk that changes in eye measurements may be wrongly attributed to mental effort when they are actually caused by visual stimuli.

This thesis, therefore, not only examines whether cognitive load can be measured but also how disturbing influences from the VR environment can be excluded or controlled, and whether it is possible to find hiccups within an assembly process. This is done through two consecutive studies: A preliminary analysis of pupil influences and a realistic assembly task with varying assistance levels. An attempt is made to develop and validate a reliable measurement method. The ultimate goal is to contribute to the development of more effective and adaptive VR training environments.

Related Work

Measuring cognitive load is a widely discussed topic in the fields of human factors, educational technology, and human-computer interaction. In recent years, there has been growing interest in using VR environments as training platforms, partly because they allow for low-risk simulations of dangerous or costly work situations. Within this context, research is also being conducted into how effective VR is as a training tool and how mental effort during such simulations can be measured.

Various studies have shown that pupil size and blinking behavior can be useful indicators of cognitive load. Biondi et al. demonstrated that participants performing a cognitively demanding task exhibited, on average, larger pupils and higher blink frequencies [Bio+23]. Other research also suggests that pupil measurements, known as pupillometry, provide insights into the level of mental effort. However, some authors point out that external visual influences, such as brightness and color of the environment, can distort the results.

Additionally, there are studies that specifically focus on cognitive load in VR. It has been shown that training in a virtual environment, when well-designed, can lead to comparable or even better performance than training in the real world. Nevertheless, measuring cognitive

load in VR remains challenging, especially due to the dynamic visual characteristics of these environments.

Finally, the literature emphasizes that both overload and underload can negatively impact performance. A well-designed training should therefore be aligned with the appropriate level of mental challenge. These insights formed the basis for the approach in this thesis, which first examined disruptive visual factors and then implemented a VR assembly task with varying levels of assistance to measure the impact on cognitive load.

Explorational Study: Influences on Pupil Behaviour

Before examining the effect of cognitive load in a realistic virtual assembly task, an exploratory study was conducted to determine to what extent external visual factors such as brightness, depth, and color influence pupil size. Since pupil size is a crucial indicator of cognitive load, it was essential first to identify these potentially disruptive factors.

For this exploratory study, a series of experiments was set up in a controlled VR environment, manipulating one factor at a time. In total, twelve participants participated in these tests, conducted with a Varjo XR-3 headset, enabling precise eye measurements.

The first test investigated the effect of brightness on pupil size. Participants performed a cognitive task (n-back test) under two different lighting conditions: normal and disturbed lighting. Results indicated that although task difficulty influenced pupil size, an abrupt decrease in brightness did not have a significant effect. This suggests that pupil changes in this context are driven primarily by cognitive load rather than brightness variations. However, this contradicts the results of the paper from which the test is taken. Therefore, these findings should be used with caution.

The second test examined whether virtual depth affects pupil size. In an empty VR environment, cubes appeared at varying distances. By carefully matching color and brightness between the background and objects, the effect of luminance was minimized. Results showed that depth alone had no significant effect on pupil size, contrasting previous findings in less controlled settings.

The third test explored the effect of color. Participants observed a wall that changed color every few seconds. Sixteen colors were tested, including both light and dark shades. The findings demonstrated that specific colors significantly influenced pupil size.

Based on these findings, it was decided to minimize color differences that significantly affect pupil size in the subsequent assembly task. Additionally, environmental brightness was kept as constant as possible. These adjustments allowed for a more accurate measurement of cognitive load in the main study.

Assembly Task in VR: Measuring cognitive workload

After the exploratory study, a realistic VR assembly task was designed to measure cognitive load within a more complex and task-oriented context. The task involved correctly labeling an electrical cabinet, which requires accuracy and focused attention. This task was performed in a tailored virtual environment, carefully controlling previously identified visual influences, such as color. Brightness is also kept as a constant because there is no need to change it.

The study was organized with two groups, each consisting of twelve participants. Both groups performed the same task, but under different conditions. The first group, the *high-assistance group*, received full support during both the training and the assembly phases: clear visual instructions, indicators highlighting the correct component, and error feedback via a red border around the object when mistakes were made. The second group, the *low-assistance group*, received full support during the training phase but had to perform the assembly phase without

visual instructions or error detection. This approach allowed investigating whether the removal of assistance results in increased cognitive load, which could be measured.

The task consisted of placing stickers on the correct components of a preassembled electrical cabinet. A total of 27 stickers had to be correctly placed across different rails. The stickers closely resembled each other and sometimes differed only in text or shape. As a result, participants had to carefully follow the instructions and perform the task with precision. The task was designed to be long enough to gather reliable measurements, yet challenging enough to effectively induce cognitive load.

Before each phase, a calibration process was conducted. First, eye tracking was calibrated, followed by measuring pupil size under minimal luminance conditions (by looking at a black wall for ten seconds), and finally establishing a general baseline by looking at the preassembled cabinet for ten seconds. These measurements were later used to normalize changes in pupil size.

During the task, several variables were recorded. Key metrics included pupil size and blinking rate. Additionally, each participant's subjective mental workload was assessed using the NASA-TLX questionnaire, evaluating six dimensions: mental load, physical load, temporal pressure, perceived performance, effort, and frustration. The total duration of each phase and the number of errors were also recorded. This combination of objective and subjective measures allowed cognitive load to be analyzed comprehensively from multiple perspectives.

Results and Interpretation

The collected data were analyzed to determine to what extent the various measured variables (such as pupil size and blink frequency) correlate with the level of cognitive load, and how this is influenced by the level of assistance during the assembly task.

High-Assistance Group

For the group that received full assistance in both phases, it was observed that participants showed lower average pupil sizes during the second phase (the repetition) compared to the first phase, although this difference was not statistically significant. This is as expected since the task had already been performed once and the instructions remained the same. Participants also spent less time reading instructions and searching for components. These time savings were accompanied by a shorter overall task completion time. However, blink frequency did not show a significant change between the two phases, indicating that this signal may be less sensitive to small differences in cognitive load when support levels remain constant.

Low-Assistance Group

In the group that had to complete the second phase without visual support, the opposite effects were observed. Pupil size was significantly larger during the assembly phase without assistance, which may indicate an increase in cognitive load due to the removal of visual guidance. The time spent searching for the correct component increased, and the total completion time was also longer than in the first phase. Blink frequency showed a significant increase in this group. The self-reported mental workload in the NASA-TLX was also higher during the second phase.

Comparison Between Groups

When comparing both groups in the second phase, it was clear that the low-assistance group exhibited, on average, larger pupil sizes than the high-assistance group. **high-assistance participants** also decreased their subjective mental load ratings, while it increased for **low-**

assistance participants. This confirms that the removal of assistance leads to higher mental effort.

Nevertheless, the relationship between cognitive load and pupil dilation was absent at the individual participant level. Participants who reported greater changes in perceived mental workload did not necessarily show a corresponding increase in pupil dilation. However, it remains difficult to truly compare this on an individual level, as each person is different. Future research will need to verify this link in more depth in order to determine whether cognitive load can be accurately measured in complex VR assembly tasks.

In summary, the results show that pupil size is a promising measure for detecting changes in cognitive load within VR, especially when visual disturbances are minimized. Blink frequency appears to be less robust, and subjective measures remain valuable for complementing and validating objective data.

Conclusion

This thesis aimed to examine to what extent cognitive load can be accurately measured during complex assembly tasks in a VR environment, using eye-based measurements such as pupil size and blink frequency. A phased approach was followed, starting with an exploratory study on visual influences, followed by a practice-oriented assembly task. The goal was to develop an accurate cognitive load measurement system using eye metrics.

The results show that pupil size can be a valuable indicator of cognitive load in a controlled VR environment. Especially when disruptive factors such as color, brightness, and virtual depth are minimized. It is shown that pupil dilation did not significantly change within the group that received the same level of assistance twice, while increasing significantly for the group that did the second phase with a lower form of assistance. Blink frequency did not prove to be a useful measure in this context.

Nevertheless, the relation between cognitive workload and pupil dilation was missing at the participant level. Participants with more perceived mental demand changes were not directly paired with a higher pupil dilation increase. However, it remains difficult to really compare at this level since all people are different. Future studies will need to verify this link on a deeper level to really be able to state whether it is possible to measure cognitive workload accurately in complex VR assembly tasks.

However, since a significant pupil dilation increase was found when assistance levels dropped, and the dilation did not change when the assistance level did not change, it shows that the adaptations done to the virtual environment and the baseline measurements were sufficient in omitting other pupil dilation influences. Thus, carefully controlling colors, brightness, and having a good baseline is key to cognitive load measurements in VR.

The final question, whether it is possible to find hiccups within an assembly process, can not be answered. There is not enough data on this to verify whether it is true or not. However, there is one example that might hint at its possibility, but more research will be required on this topic.

This thesis offers a valuable contribution to the field of cognitive load measurement in VR and highlights the importance of controlled visual environments, multi-modal measurement strategies, and task design tailored to cognitive assessment.

Based on these findings, there is a strong indication of the possibility of accurately measuring cognitive load in more complex virtual environments. However, there are still some uncertainties, which is why **further validation is required to confidently state that measuring pupil dilation in complex VR-environments using eye measurements can be accurately done.**.

Samenvatting

Inleiding

In dit onderzoek wordt nagegaan hoe de cognitieve belasting van personen accuraat gemeten kan worden tijdens het uitvoeren van complexe assemblagetaken in een virtuele realiteit (VR)-omgeving. Met de opkomst van VR als trainingsinstrument in industriële contexten, zoals bij het simuleren van werksituaties in olieplatformen of elektriciteitscentrales, groeit ook de nood aan betrouwbare methoden om mentale belasting te meten. Een correcte inschatting van die belasting is essentieel om te kunnen oordelen over de effectiviteit van training en om overbelasting of onderbelasting te vermijden.

Eerdere studies tonen aan dat oogmetingen, zoals pupilgrootte en knipperfrequentie, sterke indicatoren kunnen zijn van cognitieve load. Omdat veel moderne VR-headsets over ingebouwde oogtracking beschikken, biedt dit een unieke kans om deze metingen op een natuurlijke manier te integreren in virtuele simulaties. Toch is het meten van cognitieve belasting via oogdata niet zonder uitdagingen. Factoren zoals helderheid van het scherm, de kleuren van objecten of de virtuele diepte van elementen kunnen een aanzienlijke invloed uitoefenen op pupilgedrag. Daardoor bestaat het risico dat veranderingen in oogmetingen foutief worden toegeschreven aan mentale inspanning, terwijl ze in werkelijkheid door visuele stimuli worden veroorzaakt.

Deze thesis onderzoekt daarom niet alleen of cognitieve belasting kan worden gemeten, maar ook hoe storende invloeden uit de VR-omgeving kunnen worden uitgesloten of gecontroleerd, en of het mogelijk is om eventuele struikelblokken te kunnen identificeren via deze metingen. Dit wordt onderzocht door middel van twee opeenvolgende studies, een verkennende analyse van pupilinvloeden en een realistische assemblagetaak met variërende assistentieniveaus – wordt getracht een betrouwbare meetmethode te ontwikkelen én te toetsen. Het uiteindelijke doel is om bij te dragen aan de ontwikkeling van effectievere en adaptieve VR-trainingsomgevingen.

Gerelateerd Werk

Het meten van cognitieve belasting is een veelbesproken onderwerp binnen de domeinen van menselijke factoren, onderwijstechnologie en human-computer interaction. In recente jaren is er toenemende aandacht gekomen voor het gebruik van VR-omgevingen als trainingsplatform, onder meer omdat ze risicoarme simulaties mogelijk maken van gevaarlijke of kostbare werksituaties. Binnen dit kader wordt ook onderzocht in welke mate VR effectief is als trainingsmiddel en hoe mentale inspanning tijdens dergelijke simulaties gemeten kan worden.

Verschillende studies hebben aangetoond dat pupilgrootte en knippergedrag bruikbare indicatoren kunnen zijn van cognitieve belasting. Biondi et al. toonden aan dat deelnemers die een cognitief uitdagende taak uitvoerden, gemiddeld grotere pupillen en hogere knipperfrequenties vertoonden [Bio+23]. Ook andere onderzoeken suggereren dat pupilmetingen, ook wel pupillometrie genoemd, inzicht geven in de mate van mentale inspanning. Toch wijzen sommige auteurs erop dat externe visuele invloeden, zoals helderheid en kleur van de omgeving, een vertekend beeld kunnen geven.

Daarnaast zijn er studies die zich specifiek richten op cognitieve belasting in VR. Zo werd aangetoond dat training in een virtuele omgeving, mits goed ontworpen, kan leiden tot vergelijkbare of zelfs betere prestaties dan training in de echte wereld. Toch blijft het meten van cognitieve belasting in VR uitdagend, vooral door de dynamische visuele eigenschappen van deze omgevingen.

Tot slot wordt in de literatuur benadrukt dat over- en onderbelasting beide nadelige effecten kunnen hebben op prestaties. Een goed ontworpen training moet daarom afgestemd zijn op het juiste niveau van mentale uitdaging. Deze inzichten vormden de basis voor de aanpak in deze thesis, waarbij eerst storende visuele factoren werden onderzocht en vervolgens een VR-assemblagetaak werd opgezet met verschillende niveaus van assistentie om de impact op cognitieve belasting te meten.

Verkennende Studie: Invloeden op Pupilgedrag

Voordat het effect van cognitieve belasting in een realistische virtuele assemblagetaak onderzocht kon worden, werd een verkennende studie uitgevoerd om te bepalen in welke mate externe visuele factoren zoals helderheid, diepte en kleur de pupilgrootte beïnvloeden. Omdat pupilgrootte een belangrijke indicator is voor cognitieve belasting, was het essentieel om eerst deze mogelijke verstorende factoren in kaart te brengen.

Voor deze studie werd een reeks experimenten opgezet in een gecontroleerde VR-omgeving, waarbij telkens één factor werd gemanipuleerd. In totaal namen twaalf deelnemers deel aan deze tests, die uitgevoerd werden met een Varjo XR-3 headset, die nauwkeurige oogmetingen mogelijk maakt.

De eerste test onderzocht het effect van helderheid op pupilgrootte. Deelnemers voerden een cognitieve taak uit (n-back test) onder twee verschillende lichtcondities: normaal en verstoord licht. De resultaten toonden aan dat, hoewel de taakzwaarte invloed had op pupilgrootte, een abrupte verlaging van de helderheid geen significant effect veroorzaakte. Dit suggereert dat pupilveranderingen in deze context eerder door cognitieve belasting dan door helderheid veroorzaakt worden. Dit contradict echter de paper waar de taak uit komt en dus moet hier toch voorzichtig mee omgegaan worden.

In de tweede test werd nagegaan of virtuele diepte invloed heeft op de pupil. In een lege VR-omgeving verschenen kubussen op verschillende afstanden. Door gebruik te maken van kleur- en helderheidsmatching tussen achtergrond en objecten, werd het effect van lichtinval geminimaliseerd. Uit de resultaten bleek dat diepte op zichzelf geen significante invloed uitoefende op pupilgrootte, wat in contrast staat met eerdere bevindingen in minder gecontroleerde omgevingen.

De derde test onderzocht het effect van kleur. Deelnemers keken naar een muur die om de paar seconden van kleur veranderde. Zestien kleuren werden getest, met zowel lichte als donkere tinten. De resultaten toonden aan dat bepaalde kleuren een duidelijk significant effect hadden op pupilgrootte.

Op basis van deze bevindingen werd besloten om in de latere assemblagetaak kleurverschillen met significante pupilverandering te minimaluseren. Ook werd de helderheid van de omgeving zo constant mogelijk gehouden. Deze aanpassingen maakten het mogelijk om in de hoofdstudie de cognitieve belasting zo zuiver mogelijk te meten.

Assemblagetaak in VR: Meting van Cognitieve Belasting

Na de verkennende studie werd een bestaande realistische VR-assemblagetaak aangepast om cognitieve belasting te meten in een complexere en meer taakgerichte context. De taak bestond uit het correct labelen van een elektrische stroomkast, een opdracht die nauwkeurigheid en aandacht vereiste. Deze taak werd uitgevoerd in een aangepaste virtuele omgeving waarin de

eerder geïdentificeerde visuele invloeden (zoals kleur) gecontroleerd werden. Ook helderheid van de scene blijft hier ingewijzgd aangezien er geen reden is om dit aan te passen.

De studie werd opgezet met twee groepen van elk twaalf deelnemers. Beide groepen voerden dezelfde taak uit, maar onder verschillende omstandigheden. De eerste groep, de *high-assistance groep*, kreeg tijdens zowel de trainings- als de assemblagefase volledige ondersteuning: duidelijke visuele instructies, aanduiding van het juiste onderdeel, en bevestiging wanneer een fout werd gemaakt via een rode rand rond het object. De tweede groep, de *low-assistance groep*, kreeg tijdens de trainingsfase nog volledige hulp, maar moest in de assemblagefase zonder visuele instructies of foutdetectie werken. Op deze manier kon onderzocht worden of het wegvallen van hulp leidt tot een verhoogde cognitieve belasting, wat dan gemeten kan worden.

De taak bestond uit het plaatsen van stickers op het juiste onderdeel van een vooraf gemonteerde elektrische kast. In totaal moesten 27 stickers correct geplaatst worden, verdeeld over verschillende rails. De stickers leken sterk op elkaar en verschilden soms enkel in tekst of vorm. Hierdoor moesten deelnemers nauwgezet de instructies volgen en zorgvuldig te werk gaan. De taak was zo ontworpen dat ze voldoende lang duurde om betrouwbare metingen te verzamelen, maar ook uitdagend genoeg was om cognitieve belasting uit te lokken.

Voorafgaand aan elke fase werd een kalibratieproces uitgevoerd. Eerst werd de oogtracking gekalibreerd, gevolgd door een meting van de pupilgrootte bij minimale luminantie (door tien seconden naar een zwarte muur te kijken) en een algemene baseline door tien seconden naar de gemonteerde kast te kijken. Deze gegevens werden later gebruikt om veranderingen in pupilgrootte te normaliseren.

Tijdens de taak werden meerdere variabelen geregistreerd. De belangrijkste waren pupilgrootte, knipperfrequentie, tijd gespendeerd aan het bekijken van instructies, en zoektijd naar het juiste onderdeel. Daarnaast werd per deelnemer het zelf ervaren mentaal werktempo vastgelegd via de NASA-TLX vragenlijst, die zes aspecten beoordeelt: mentale belasting, fysieke belasting, tijdsdruk, eigen prestaties, inspanning en frustratie. Ook de totale duurtijd van elke fase en het aantal fouten werd bijgehouden. Deze combinatie van objectieve en subjectieve metingen liet toe om cognitieve belasting vanuit meerdere perspectieven te analyseren.

Resultaten en Interpretatie

De verzamelde gegevens werden geanalyseerd om te bepalen in hoeverre de verschillende meetwaarden (zoals pupilgrootte en knipperfrequentie) samenhangen met het niveau van cognitieve belasting, en hoe dit beïnvloed wordt door het assistentieniveau tijdens de assemblagetaak.

High-Assistance groep

Voor de groep met volledige hulp in beide fases bleek dat deelnemers tijdens de tweede fase (de herhaling) gemiddeld lagere pupilwaarden vertoonden dan in de eerste fase, ook al was dit niet significant. Dit is logisch aangezien de taak al eerder werd uitgevoerd en de instructies identiek bleven. Deelnemers besteedden ook minder tijd aan het lezen van instructies en het zoeken naar onderdelen. Deze tijdswinsten gingen gepaard met een kortere totale voltooiingstijd van de taak. De knipperfrequentie vertoonde echter ook geen significante verandering tussen beide fases, wat aangeeft dat dit signaal mogelijk minder gevoelig is voor kleine verschillen in belasting bij gelijkblijvende ondersteuning.

Low-Assistance groep

Bij de groep die tijdens de tweede fase zonder visuele ondersteuning moest werken, werden omgekeerde effecten vastgesteld. De pupilgrootte was gemiddeld groter tijdens de assemblagefase zonder hulp, wat kan wijzen op een toename van cognitieve belasting door het wegvallen

van de visuele ondersteuning. De zoektijd naar het juiste onderdeel nam toe, en ook de totale voltooiingstijd was langer dan tijdens de eerste fase. De knipperfrequentie vertoonde bij deze groep een significante stijging. De zelfgerapporteerde mentale belasting in de NASA-TLX was eveneens hoger bij de tweede fase.

Vergelijking tussen groepen

Bij vergelijking tussen beide groepen in de tweede fase bleek duidelijk dat de low-assistance groep gemiddeld een grotere pupilgrootte vertoonde dan de high-assistance groep. **high-assistance deelnemers** gaven in de 2de fase ook lagere mental load scores, terwijl **low-assistance** deelnemers hier juist hogere mental load scores gaven. Dit bevestigt dat het wegnemen van ondersteuning leidt tot hogere mentale inspanning.

Desondanks ontbrak de relatie tussen cognitieve belasting en pupilverwijding op deelnemerniveau. Deelnemers die meer verandering in ervaren mentale belasting rapporteerden, vertoonden niet noodzakelijkerwijs een grotere toename in pupilverwijding. Het blijft echter moeilijk om dit op individueel niveau echt te vergelijken, aangezien ieder persoon verschillend is. Toekomstig onderzoek zal deze link op een dieper niveau moeten verifiëren om echt te kunnen stellen of cognitieve belasting accuraat meetbaar is in complexe VR-montagetaken.

In het algemeen tonen de resultaten aan dat pupilgrootte een veelbelovende maat is voor het detecteren van veranderingen in cognitieve belasting binnen VR, vooral wanneer visuele storingen geminimaliseerd worden. Knipperfrequentie lijkt minder robuust, en subjectieve metingen blijven waardevol ter aanvulling en validatie van objectieve data.

Conclusie

Deze masterproef had als doel om te onderzoeken in welke mate cognitieve belasting accuraat gemeten kan worden tijdens complexe assemblagetaken in een VR-omgeving, met behulp van op het oog gebaseerde metingen zoals pupilgrootte en knipperfrequentie. Er werd een gefaseerde aanpak gevolgd, beginnend met een verkennende studie naar visuele invloeden, gevolgd door een praktijkgerichte assemblagetaak. Het doel was om een nauwkeurig meetsysteem voor cognitieve belasting te ontwikkelen op basis van oogmetingen.

De resultaten tonen aan dat pupilgrootte een waardevolle indicator kan zijn van cognitieve belasting in een gecontroleerde VR-omgeving, vooral wanneer storende factoren zoals kleur, helderheid en virtuele diepte geminimaliseerd worden. Er is ondervonden dat de pupilgrootte niet significant veranderde voor deelnemers wiens assistentie level hetzelfde bleek terwijl pupilgrootte significant vergrootte bij die groep die in de 2de fase een lagere vorm van assistentie kreeg. Knipperfrequentie bleek in deze context geen bruikbare maat te zijn.

Toch ontbrak de relatie tussen cognitieve belasting en pupilverwijding op individueel niveau. Deelnemers die meer veranderingen in mentale belasting rapporteerden, vertoonden niet automatisch een grotere toename in pupilverwijding. Het blijft echter moeilijk om dit op dit niveau goed te vergelijken, aangezien elk individu verschillend is. Toekomstig onderzoek zal deze link op een dieper niveau moeten verifiëren om echt te kunnen bepalen of cognitieve belasting accuraat meetbaar is in complexe VR-assemblagetaken.

Aangezien er echter een significante toename in pupilverwijding werd vastgesteld bij afname van assistentie, en er geen verandering was wanneer het assistentieniveau gelijk bleef, toont dit aan dat de aanpassingen aan de virtuele omgeving en de basismetingen voldoende waren om andere invloeden op pupilverwijding uit te sluiten. Het zorgvuldig controleren van kleuren, helderheid en het hanteren van een goede basismeting zijn dus cruciaal voor het meten van cognitieve belasting in VR.

De uiteindelijke vraag, of het mogelijk is om haperingen binnen een assemblageproces op te sporen, kan niet beantwoord worden. Er is onvoldoende data om te verifiëren of dit al dan

niet het geval is. Er is echter één voorbeeld dat op deze mogelijkheid zou kunnen wijzen, maar verder onderzoek is nodig op dit vlak.

Deze masterproef levert een waardevolle bijdrage aan het vakgebied van cognitieve belasting-meting in VR en onderstreept het belang van gecontroleerde visuele omgevingen, multimodale meetstrategieën en een taakontwerp dat afgestemd is op cognitieve evaluatie.

Op basis van wat er ondervonden is, is er een zeer sterke indicatie dat het mogelijk is om cognitieve belasting te meten in meer complexe virtuele omgevingen. Echter, zijn er nog wat onzekerheden en dus **is verdere validatie nodig om met zekerheid te kunnen stellen dat cognitieve lading accuraat kan gemeten worden in complexe VR-omgevinden via oog metingen.**

Contents

Chapter 1

Introduction

In the modern world, changes and technological improvements occur rapidly. Everything is constantly changing with respect to speed, agility, accuracy, and improvement. With all these changes, it becomes increasingly difficult for companies to keep their employees up-to-date with the latest technology. For this, employees must undergo training, which is not always easy. A company might have jobs where training employees on site is impossible due to safety hazards, such as oil rigs or power plants. This results in employees who lack "hands-on" experience with the technical working environment [Sul+20; Sur+20]. Developing training facilities where all the machinery is placed for training would be possible, but this is a serious investment since most machinery is expensive. Extensive facilities would have to be built, which also comes at a significant cost. Developing simulations of the required machinery could reduce these costs. For example, Akulov *et al.* created a training simulator for a crane operator. The simulator is equipped with the physical interface of an actual crane and displays the panorama view on a screen in front of the trainee [Aku+23]. This allows trainees to fully immerse themselves in the experience.

According to Shaikh *et al.* assembly work often involves concurrent performance, which is both physical and mental. Assembly tasks mostly involve awkward postures, manipulating components with your hands, memorizing defined procedures and part numbers, rapid information processing, and decision-making. This all is done while there is a deadline for the assembly, which puts even more pressure on the employee [Sha+12]. Employers have high expectations of their employees, even though they may not always be aware of this. Besides, it is also hard to measure an employee's mental state while working on an assembly.

Virtual reality (VR) could be used to better immerse trainees into the work environment. It is increasingly used daily, mainly as a popular game tool, because it allows you to immerse yourself fully into the game's world. VR can also give new opportunities for business use. For example, VR can train employees to do specific tasks, like the ones done on an oil rig or power plant. Thanks to VR, companies can let their employees train inside safe environments where their mistakes do not have consequences. Even tasks that typically cost a lot of resources or have high risk can be repeated in the virtual environment for training. [GSS24]

So VR gives employees the ability to train in a virtual environment. This leads to lower risks because employees are using virtual tools. Besides, employees can be easily put in an isolated environment. Finally, many headsets have built-in eye tracking. This means that every eye movement of the employees can be tracked and analyzed, giving the ability to see what the trainee looked at, how the trainee's blinking pattern was, and even the pupil dilation of the trainee during training. These eye trackers also give the ability to measure the cognitive load of a trainee while doing the training session. Measuring different aspects of human eyes, such as pupil dilation and blinking rate, has proven to be an accurate measurement of cognitive load [EHR21; PS03; SS24]. Some high-end VR headsets come with eye-tracking, which enables

cognitive load measurement. It can measure pupil dilation in real time, report what the user looks at inside the virtual environment (gaze patterns), and count how much the user blinks. Pupil dilation [Sev+22; NM24] and blinking rate [Gur+24] is an interesting metric to measure since it is associated with one's performance.

Cognitive load is the amount of working memory that a human brain is using at any point of time. This working memory has a limited capacity, which can be overloaded, resulting in a reduced performance [CB23]. It is shown that for aviation pilots, cognitive load directly affects their performance [Yu+23]. Since oil rigs and power plants also have complex machinery to work with, this will also apply for these employees. Cognitive load can be measured using different methods like EEG and GSR sensors [Ahm+23].

Since virtual training seems promising for companies while VR headsets also allow for eye-tracking possibilities, this thesis studies whether cognitive load can be accurately measured when performing realistic assembly tasks within a virtual environment. It also investigates how these virtual environments must be designed to measure the cognitive load without any interference from external factors like the brightness of the VR screen or the fact that everything the user sees is virtual. Screen brightness is an important point here because Pignoni *et al.* mention that the magnitude of light's change on the pupil diameter can be ten times the measurable influence that the cognitive load has. [PKV21]. Anyhow, measuring cognitive load in assembly training can be very beneficial because it allows for training parameters to be adapted based on how high or low the cognitive load is measured to be. Besides, when developing new methods for assembly, the new processes can be developed virtually and checked for any cognitive load bottlenecks before actually implementing the assembly process into the company.

Measuring cognitive load in virtual environments for complex tasks is a critical challenge in human-computer interaction. Some factors influence the measurements, which must be eliminated as much as possible. This brings us to the **research questions**:

- Can the cognitive load be accurately captured in virtual environments containing complex assembly tasks via eye measurements?
- How can interfering factors be eliminated
- Is it possible to accurately find the hiccups within an assembly process

Addressing these questions can provide valuable insight into the interplay between virtual environment design and cognitive load assessment.

In Unity, a realistic assembly task is created for an electric cabinet to investigate the optimal design and the accurate measurement of the cognitive load. This task is used in a user study, where methods are used to extract data on indications of cognitive load, such as pupil dilation, blinking rate, and gaze patterns. The user study is described in chapter 4. The results of this user study are shown in the chapter 5.

Chapter 2

Related Work

There has already been some research on training people and measuring cognitive workload inside virtual environments. This chapter reviews past work related to studies involving cognitive load measurements.

Li *et al.* [Li+22] determined if VR environments influence learning ability. Participants were asked to play Tetris on a screen and inside a VR environment. They were asked to play twice in a real-world setting and twice in a VR environment. One of these times per environment is with a Detection Response Task (DRT). Participants were fitted with a vibrating device around their necks, and whenever they felt a vibration, they had to press a pedal placed on the floor. The paper states no significant differences in cognitive load between real-world and virtual training. The VR environment also did not hinder the task performance. Oren *et al.* [Ore+12] performed a similar study to investigate the difference in the training of assembly tasks between real-world and virtual training environments. After doing an entrance survey and watching a tutorial video on how to use the devices, participants were trained to solve a wooden burr puzzle by looking at a color-coded example paper and either actual blocks or blocks in a virtual environment. Eventually, both the real-world and virtual trainees had to complete the burr puzzle using real blocks without the color codes. The virtual group took 3.63 times longer to finish the training but completed the puzzle 1.78 times faster than the real-world group. Also, all participants from the virtual group could finish the puzzle, while the real-world group had two participants who could not finish it. This means virtual assembly might benefit training employees for assembly tasks. Both papers have shown that training employees in a virtual environment might be beneficial. This enforces the positive impact that cognitive workload measurements could bring. If being in VR can simulate the real world scenarios this well, the cognitive load measurements found might relate to the real world as well, meaning the findings in VR are also meaningful for real-world tasks.

Biondi *et al.* [Bio+23] studied if it is possible to measure cognitive load using pupil dilation and blinking rate. Participants had to perform a physical task in combination with a cognitively demanding task. For the physical task, participants had to pull levers and insert hoses. There is an extended version and a short version for the physical task. For the cognitive task, there was an easy one and a hard one. The easy scenario meant users had to perform the physical task with no cognitive task, and the hard scenario meant participants had to do the short or long physical task alongside a 2-back test. The physical task was built to act like a common task in the automotive manufacturing industry, while the cognitive task is more about simulating cognitive load comparable to listening to radio or making phone calls. The paper states that both pupil size and blinking rate were sensitive to increased cognitive load while performing the 2-back task. A greater normalized pupil size was also observed during the high cognitive task. Eckert *et al.* [EHR21] did a similar study on how cognitive load can be measured through pupil dilation with off-the-shelf VR headsets without the need for steady lighting conditions. A

key challenge they addressed was the impact of the pupillary light reflex, which can obscure the cognitive load signal. To overcome this, they developed a method to correct for the pupil dilation caused by the pupillary light reflex. Participants must perform an n-back task ranging from 1-back to 3-back in multiple light conditions. The normal condition means there is just standard lighting, and the corrupt lighting implies that after the 10th ball, the overall scene brightness will drop by 50 percent, thus introducing "unpredicted" lighting. The study revealed significant effects on pupil dilation for both the task and the lighting conditions before the correction. After the correction, the task level still showed significant differences, while lighting was no longer a main effect. Thus, according to this study, an off-the-shelf VR headset with eye-tracking will suffice to measure cognitive load. Lee *et al.* [Lee+24] also studied whether pupillometry can be used to measure cognitive workload while performing tasks of different toughness correctly. For this test, two tasks in the healthcare industry were created in a virtual environment, both having different difficulty levels. The easy task just had the instruction of "Observe and report: What is the homecare provider doing in the scenario". A more detailed observation was required for the difficult task, with multiple questions left to answer. According to the study, the difficult task resulted in a higher cognitive load measurement in pupil dilation. According to this study, cognitive load can be measured using pupillometry. The paper also mentions that a traditional way to counter pupillary light reflex is to get a baseline recording in the same lighting conditions as the test. However, when using real-world VR scenarios, using fixed lighting conditions is not always practical. The paper also suggests an alternative: developing baseline formulas to compute pupil dilation based on the changing luminance level.

Garcia *et al.* [GVM22] studied identifying phases where workload under-arousal and over-arousal may be present. Under-arousal can occur when a task is too easy, resulting in the task performer getting bored and thus performing poorly. Over-arousal can occur when there is too much information to process, resulting in the task performer's exhaustion and fatigue. The study was conducted by having participants complete a NASA Task Load Index (TLX), a multi-dimensional rating procedure for overall workload scores. It is based on the weighted average of six subscales: Mental Demand, Physical Demand, Temporal Demand, Own Performance, Effort, and Frustration. Participants were asked to fill in the NASA-TLX form online. After the questionnaire, an evaluation was conducted by having the participants work with these automated drilling processes. This is done for the automated drilling process with a robotics platform and the automated drilling process without. The results of this paper state that the automated drilling process using the robotics platform resulted in average workloads that are in the higher range of the 0-20 scale for most of the drilling process. This means that workers may become exhausted and fatigued during the drilling process. On the other hand, the automated drilling process without the robotics platform had an average workload in the lower range of the 0-20 scale. In this study, the indications done via the NASA-TLX can possibly be found via the eye measurement data, which would have verified whether the participants' real cognitive demand supported their claims. By studying participants' cognitive abilities in a serious game environment, Jerčič *et al.* [JSL17] investigated how a high cognitive load can overshadow the physiological arousal effect. To measure cognitive load, pupil dilation and heart rate were measured. The serious game was based on an auction game. Participants had to decide by calculating a mean value from three price estimations and then decide to sell or buy at the correct price. Price estimations were linked to the physiological arousal level. This was done so that the more the price deviated from the correct price, the more aroused the participant was. This results in lower physiological arousal when the variance of price estimations becomes closer to the correct price, making the buy/sell decision easier. The task became more challenging because the decision period was reduced, forcing quick decision-making. The study suggests that pupil dilation shows the moment when cognitive load overshadows the physiological arousal effect. These findings suggest that pupil dilation can help measure cognitive workload if combined with another metric, like the heart rate. This is because pupil dilation changes swiftly and constantly. Also, Abdurrahman *et al.* [Abd+22]. studied how cognitive workload can affect task performance by having students drive from a start point to a destination using landmarks. The cognitive workload was measured by

measuring the participants' pupil dilation and heart rate. Participants had to navigate either an easy road with essential landmarks and an easy road with non-essential landmarks or a hard road with essential and non-essential landmarks. Essential landmarks mean that the landmarks were located at intersections. This means the landmarks were right where you had to make a decision, thus resulting in more straightforward navigation. Non-essential landmarks were located on the sides of the routes. To make the hard routes even more challenging, the routes had five traffic lights, turns, and intersections, while the easy ones only had 3. The study has shown that the easier routes were completed much faster than the more complex routes. It is also shown that non-essential landmarks on complex routes increased psychophysiological activity, meaning pupil dilation and heart rate went up. This happens because more cognitive resources were demanded while navigating this challenging route. Also, when the cognitive demand increased, the participants started making more mistakes, which meant the cognitive load affected the performance.

Another interesting study by Iskander *et al.* [Isk+19], is done to determine if eyes respond to virtual depth. Since complex tasks mostly consist of moving objects, there will most likely be objects at different depths or even objects changing in depth. Therefore, it is crucial to know how pupils respond to virtual depth. Object distance and pupil size positively correlate in a real-world setting, meaning if the distance between an object and your eyes grows, your pupils also grow. The test consisted of an empty Unity world with a grid of cubes in it. The grid consisted of 9 cubes per depth. Per depth, five cubes were shown one by one for 1.5 seconds before moving to the next depth. Iskander *et al.* [Isk+19] states that the correlation in a VR environment is negative, which is the opposite of a real-world environment. This means that your pupils will shrink when an object goes further away from your point of view. This study is important to keep in mind since the VR environment will be 3D, meaning objects have distances, which might also influence pupil dilation according to this study.

Much of the related work involves measuring cognitive workload by pupil dilation. Measuring pupils' size change is a very popular way to indicate the cognitive workload. Because of this, this study also uses pupil dilation as an indication of cognitive workload as the base measurement. One of the studies also mentions a NASA-TLX to capture subjective perception. This is also interesting to implement into this study as well.

Chapter 3

Exploring Pupil Influences

Before implementing cognitive load measurements in more complex virtual environments, it is essential to know which cues affect the pupil of the eyes, since pupil dilation will be the primary measurement and might react differently to different cues within a VR environment. Therefore, this chapter will dive into some of these aspects and investigate how pupil dilation is affected by certain factors in VR.

The user study of two papers was replicated to examine which aspects affect pupil dilation. These studies are related to the work of Eckert *et al.* [EHR21] and Iskander *et al.* [Isk+19]. Since environments consist of multiple colors, it is helpful to know how pupil dilation is affected by different colors, which is why a color test was performed. The order in which the tests were performed was balanced using a Latin square. Twelve participants conducted the test, of which eleven were male and one was female. The average age is 23 years, and all but one participant studies "Informatica" at Hasselt University. During these tests, pupil size, gaze direction, calculated luminance, and blinking rate were recorded throughout the test. A mental load score is also filled in after each experiment is complete.

3.1 Material

To conduct these tests, participants were given a Varjo XR-3 mixed-reality headset. This headset comes with excellent hand-tracking capabilities and can also measure different metrics from the eyes, such as pupil dilation, gaze direction, and blinks, thanks to its eye-tracking capabilities. The development of the environment was done using Unity. The scripts that handle and log the measurements are based on existing code from Lila Boukabous. After completing every experiment, participants were asked to rate the perceived mental load score from the NASA-TLX.

In this study, three different experiments were conducted using a Latin-square configuration. The first test is named the "Brightness Test". Participants perform an n-back test in two different lighting conditions in this test. The second task is named the "Depth Test". This test is conducted to investigate whether pupils react to the depth of virtual objects. The last test is called the "Color test". This test is conducted to examine how the pupils react to different colors.

3.2 Brightness Test

The first test replicates the study by Eckert *et al.* [EHR21]. Users were put in a condo-like VR environment with a table containing two bins for this test, shown in Figure 3.1. The left bin had the n-value with backed displayed on it (1-backed, 2-backed, 3-backed), and the right one had not n-backed displayed (not 1-backed, not 2-backed, not 3-backed).

Before starting the task, users were shown a paper that provided textual and visual information about how the n-back test works. After reading this and confirming they understood it, they were given the headset to put on and start the trial.

As stated in the paper, before participants could start the test, they had to perform a calibration process. This process consisted of a sequence of colored balls spawning one by one on the table between the two bins, where the user had to look at them. Eight different colors were used, and for each ball, the overall brightness of the scene jumped from 10% to 30% and then to 50%, with each 3 seconds in between. In the paper, this was done before every round. Due to this being a short confirmation, the calibration process was only done at the start of one trial, which goes from 1-back to 3-back. The goal of this calibration is to get reference data to calculate the pupil change in Z-score.

After the calibration, users had to use the hand-tracking functionality to grab a ball and put it in the right bin. Every test had the same sequence of 1-back, 2-back, or 3-back. This trial was then done twice, once with normal and once with corrupted lighting conditions. The corrupted lighting conditions mean that after the 10th ball has been sorted, the overall brightness will drop by 50%, which should influence the pupil dilation. This is done to comply with the work of Eckert *et al.* [EHR21]. Each n-back consisted of 20 balls with the same sequence for each participant.

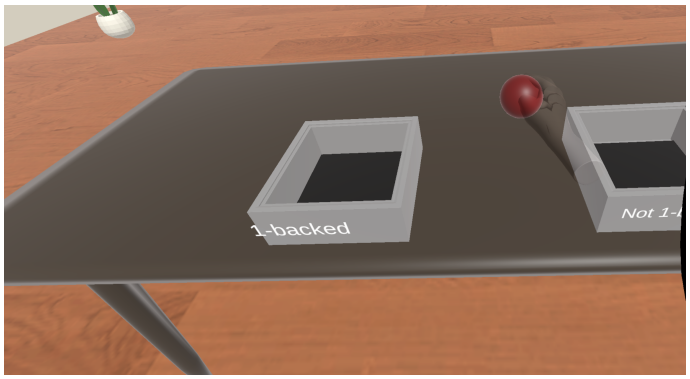


Figure 3.1: Image showing the brightness test (performed via N-back test). Two labeled bins can be seen to sort the balls. The hands are made transparent to ensure they do not emit much light, contrasting with the environment.

3.3 Depth Test

The second test was performed to verify whether the virtual depth of an object has the same influence on pupil dilation as in a real-world setting. A test was conducted to replicate the results from Iskander *et al.* [Isk+19]. According to this paper, pupil dilation and the depth of a virtual object have a negative correlation. Pupils will shrink if the distance between a virtual object and your view grows, and the pupils will grow if the distance shrinks. This means that a VR environment will have the opposite effect of a real-world setting since pupil dilation and object distance have a positive correlation [Auc18].

For this test, users are set in an empty world only consisting of a single color of blue, as seen in Figure 3.2. In this blue void, cubes will appear individually, each lasting 1.5 seconds. The cube

grid is built of five 3x3 grids of cubes. These grids have a depth of 1.5m, 1.75m, 2m, 3m, and 4m. The sequence for every participant was the same: five random cubes at 2m depth, which is the baseline, followed by 3m depth, 4m depth, back to 2m depth, then 1.75m, and lastly, 1.5m depth. At every depth, the five cubes were randomly selected without the possibility of a cube coming twice in a row at the same depth. The color of the cube was purple. Participants were instructed to keep a close eye on the cube by only moving their eyes.

The reason for the environment being a blue void and the cubes being purple is to eliminate the pupillary light response as much as possible. In the paper by Iskander *et al.*, they kept the brightness as constant as possible, but the background of the environment had a lot of contrast, which might influence pupil behavior; therefore, in this test, the background is constantly blue. The cube's color was decided based on a calculation found in the paper mentioned above by Eckert *et al.*, where they calculated the luminance of an object. First, the calculation was done on the background to get a luminance value for it. Once this value was known, the cube color was chosen so that its luminance value would closely match the luminance value of the background. This ensures that brightness will be canceled out more because if a cube is darker than its environment and comes closer, the screen will emit a less bright color, thus resulting in a pupillary response.

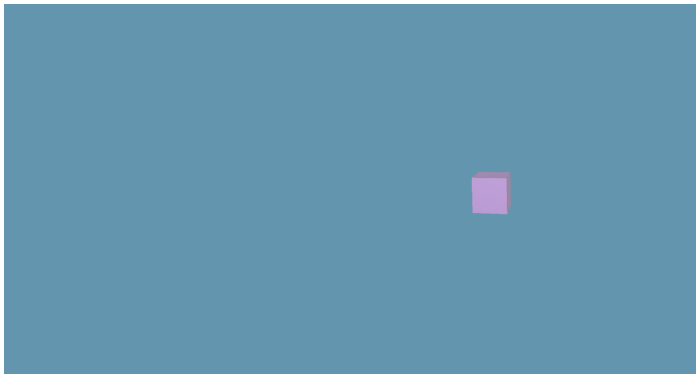


Figure 3.2: Image showing the depth test. The whole environment is blue, and purple cubes appear one by one. These colors are chosen because of their similar calculated luminance.

3.4 Color Test

The last test was conducted to see if color significantly impacts pupil dilation. Participants were put into a condo-like environment and were instructed to look at a wall, as seen in Figure 3.3. This wall would remain white for 10 seconds and then change color every 3 seconds. The sequence of colors was randomized for every participant. Sixteen colors were used, of which eight are bright and eight are dark. Every color occurs twice, but never twice in a row. The difference in pupil dilation is calculated based on the pupil's average size during the color and the 10-second white wall stare.

Figure 3.3: Image showing the color test. One of the walls inside the dojo-like building changes color.

3.5 Results

Since both papers from the replicated user studies mention their results in specific units, this study will use the same units used in their papers. This allows for a good comparison between the results of the paper and the results found in this study. For the brightness test, the units are Z-scores. Z-scores are standard scores or measures that show how much a given value deviates from the mean. In this case, the Z-index represents the number of standard deviations that the pupil's dilation deviates from the mean pupil size captured during the calibration sequence. The paper about virtual depth shows pupil dilation in percentage, which is also used for this depth test. A Shapiro-Wilk test was conducted on all the data to investigate whether the data were normally distributed. Afterward, a Friedman test was used to examine if any groups had significant differences. The groups are color for the color test, depth for the depth test, and n-back level with lighting condition for the brightness test. If there were significant differences, pairwise Wilcoxon was used for the post-hoc test since in all cases, the data were not normally distributed. The color test is not from a paper. For this test, showing the percentage was chosen because percentage change allow the data to vary more, which is helpful in making sure we get the correct results in terms of effects on pupil dilation due to colors.

3.5.1 Brightness Test

The discussed paper states that significant differences were found between the three N-back levels within the same lighting conditions. Also, significant differences exist between the results in normal and corrupted mode. For the differences in the same lighting conditions, the hypothesis is:

- H_{01}: No significant differences in pupil dilation between the N-back levels within the same lighting conditions exist.
- H_{11}: Significant differences exist in pupil dilation between the N-back levels within the same lighting conditions.

Besides, the hypothesis for the differences between the lighting conditions is

- H_{02}: The two lighting conditions show no significant differences in pupil dilation
- H_{12}: The two lighting conditions show significant differences in pupil dilation.

After replicating the test to see if these results could be replicated, significant pupil dilation differences were found between the different N-back levels, $Q(2) = 14.36, p < 0.001, W = 0.65$. A Wilcoxon post-hoc test revealed these differences are between a 1-back and a 2-back, as well as between a 1-back and a 3-back for both conditions. However, no significant differences were found between the 2-back and 3-back levels. These results can be seen in Table 3.1. Since there

is a significant difference between 1-back and 2-back, and between 1-back and 3-back, **H_{01} can be rejected**.

Although no significant differences are found between the two- and three-back tests, there is a big difference in the perceived mental load scores of the participants. Figure 3.4 shows the average ratings of experienced toughness per N-back level and per lighting conditions. It is seen that although pupil dilation did not significantly change, toughness ratings are close to being doubled. The reason for this could potentially be that when working memory is surpassed, pupil dilation can level off, as stated by El Haj *et al.* [El 24]. Figure 3.5 shows the pupil dilation for each N-back level and each lighting condition. Outliers above a Z-index of 10 and below a Z-index of -8 are removed to clarify the boxplot. This ensures that the difference can be seen more clearly.

Besides the differences between levels within the same lighting conditions, it is also possible to have differences between them. Files are created to put one n-back level together with normal and corrupted lighting values. Looking at the test results, it appears that no significant differences were found between the normal and corrupted light conditions by Friedman, for N1: $Q(1) = 1.6, p > 0.05, W = 0.07$, for N2: $Q(1) = 1.6, p > 0.05, W = 0.16$, and finally for N3: $Q(1) = 0.82, p > 0.05, W = 0.07$. This means that dropping the overall scene's brightness by 50% does not significantly impact the participant's pupil dilation, which is great for when cognitive load is to be measured. The **H_{02} hypothesis can hereby be accepted**, contradicting what the paper suggested. Therefore, it should be taken with caution.

Besides pupil dilation, the blinking rate might also be an interesting measure to help indicate cognitive load. Table 3.2 displays the differences in blinks per minute between each N-back level in both lighting conditions. 1-2 Diff means the difference between blinks per minute during the 1-back and 2-back tasks. The same applies for 1-3 Diff and 2-3 Diff. For the difference between 1-back and 2-back, we can see that across both lighting conditions, 52% of the time, the blinks per minute have increased in the 2-back test compared to the 1-back test. However, the blinks per minute decreased by one blink for the normal lighting and increased by 10 blinks for corrupted lighting on average. This is an average increase across both lighting conditions of 5.5 blinks. For the difference between 1-back and 3-back, 76% of the time, the blinks per minute are increased in the 3-back test compared to the 1-back test. Besides. This is an increase of 13 blinks in normal lighting and an increase of 11 blinks for corrupted on average, thus 12 blinks across both conditions. Finally, 76% of the participants increased their blinking rate in the 3-back test when compared to the 2-back test. Here, normal lighting increased by 15 blinks and corrupted lighting by 2 blinks per minute on average, thus resulting in an average increase of 7.5 blinks across conditions. A Friedman test is used to see if the number of blinks per minute significantly affects the different N-back levels. The Friedman test did not reveal a significant difference in blinking rates between the N-back levels for normal lighting: $Q(2) = 3.8, p > 0.05, W = 0.17$. However, corrupted lighting gave $Q(2) = 10.31, p < 0.01, W = 0.52$. Yet, the pairwise Wilcoxon test with Bonferroni correction resulted in none of the pairs actually being significantly different. This suggests that while there is an overall effect, the specific pairwise differences may be subtle, or the sample size may be insufficient to detect them individually after correction.

Test		**Normal**						**Corrupted**					
A	B	Am	Bm	W	Z	p	r	Am	Bm	W	Z	p	r
1-back	2-back	1.04	1.57	0.0	-2.93	$\mathbf{p < 0.001}$	0.96	1.48	1.79	0.0	-2.93	$\mathbf{p < 0.001}$	0.97
1-back	3-back	1.04	1.52	2.0	-2.76	$\mathbf{p < 0.01}$	0.97	1.48	2.06	9.0	-2.13	$\mathbf{p < 0.05}$	0.89
2-back	3-back	1.57	1.52	22.0	-0.98	$p > 0.05$	0.97	1.79	2.06	29.0	-0.36	$p > 0.05$	0.91

Table 3.1: Comparison of normal and corrupted p-values for different n-backs. Am means the median value of A, and Bm is the median value of B.

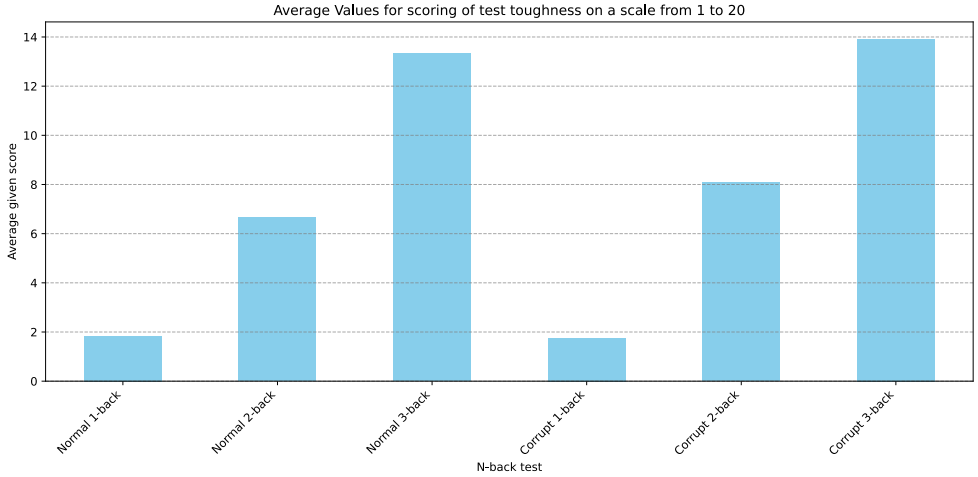


Figure 3.4: The average rated toughness felt by participants while doing a certain N-back level.

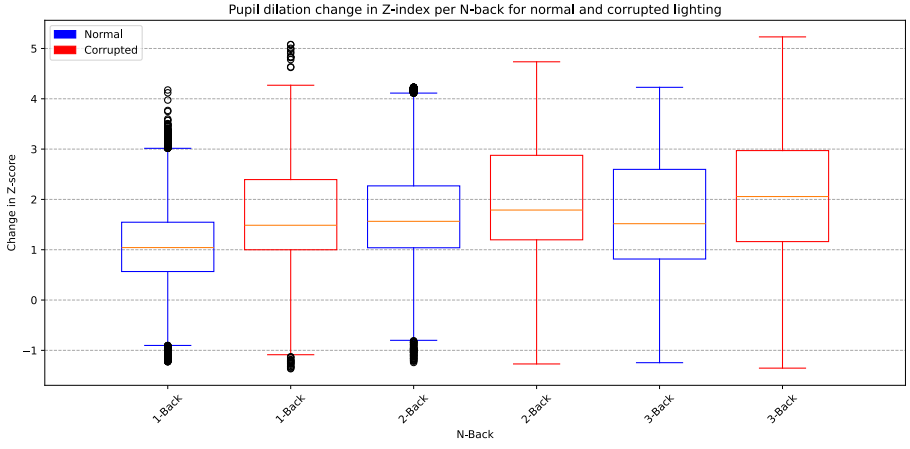


Figure 3.5: The Z-score of pupil dilation by N-back level and lighting condition. The Z-score is calculated in comparison to the mean pupil size captured during the calibration sequence. For this graph, outliers are removed.

Participant	Normal			Corrupted		
	1-2 Diff	1-3 Diff	2-3 Diff	1-2 Diff	1-3 Diff	2-3 Diff
P1	+2	-1	-3	-12	-120	-108
P2	/	/	/	/	/	/
P3	-60	+3	+63	-87	-138	-51
P4	/	/	/	+5	+7	+2
P5	+38	+45	+7	+28	+45	+17
P6	+16	+18	+2	+16	+18	+2
P7	+5	+14	+9	+7	+12	+5
P8	-2	+6	+8	+16	+47	+31
P9	-18	-24	-6	-3	+6	+9
P10	-5	+61	+66	-92	-30	+62
P11	+0	+4	+4	-3	+1	+4
P12	+12	+8	-4	+20	+28	+8
Average	-1	+13	+15	+10	+11	+2

Table 3.2: Comparison of blinks per minute between the different N-back levels in normal and corrupted lighting conditions. A positive value indicates an increase of blinks per minute in the higher N-level, while a negative value indicates a decrease in blinks per minute in the higher N-level. Values with / are 0 and are not used in calculating the average.

3.5.2 Depth Test

The discussed paper states that virtual depth negatively correlates with pupil dilation in VR environments. This gives the following hypothesis to check:

- H_{03}: There are no significant differences in pupil dilation between the depths
- H_{13}: There are significant differences in pupil dilation between the depths.

No significant differences were found between any of the depths after removing the contrasting background and giving the cubes a color similar to the background ($Q(4) = 1.53, p > 0.05, W = 0.03$). Figure 3.6 shows the pupil dilation per depth for all the participants combined. Since no significant differences were found, in contrast to the paper, the virtual depth might not have been the major influence on the pupil, but rather the contrasting background. If an object is closer, it appears bigger. This means that more of the light background is covered with a darker color, thus resulting in pupil dilation. Therefore, the closer and thus bigger the object, the bigger the pupil size. These results indicate that $\boldsymbol{H_{03}}$ **can be accepted**.

3.5.3 Color Test

Figure 3.7 shows the percentage of change in pupil dilation for all the colors. These values represent the change of the pupil's size during each color compared to the median pupil size captured during the 10-second white wall stare. Dark colors generally have a higher pupil growth than the white wall. This is because pupils tend to shrink in darker conditions [Auc18]. For this test, 16 colors were used. The dark colors are black, dark blue, dark green, dark pink, dark purple, dark red, and dark turquoise. This led to the following hypothesis:

- H_{04}: There are no significant differences between colors
- H_{14}: There are significant differences between colors.

The Friedman test showed that there are significant differences present in the test ($Q(15) = 171.84, p < 0.001, W = 0.82$). The significantly different color values can be seen in Table 3.3. The combinations not included were not significant. The light color values are blue, green, light green, yellow, orange, pink, red, and turquoise. Table 3.4 shows the significant

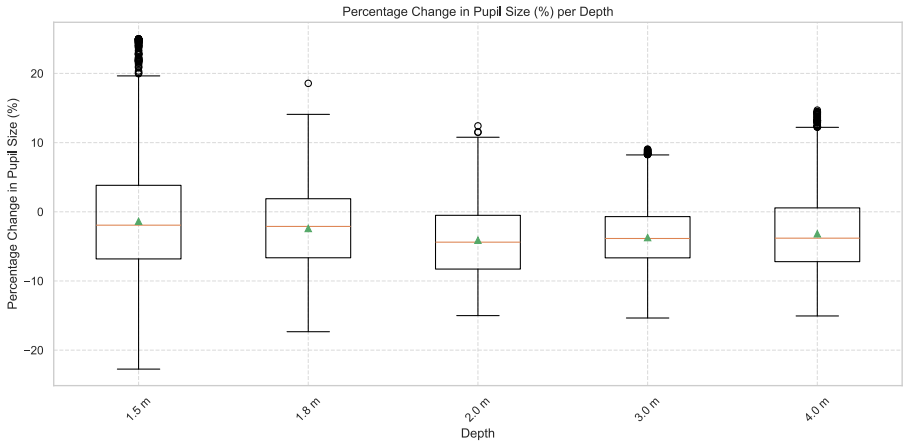


Figure 3.6: The percentage of pupil dilation by depth. This is the percentage of dilation of the pupil's size in contrast to the median pupil size captured during the calibration round, which was the first 5 blocks at a depth of 2 meters.

differences between the colors from the light group. This table also only displays the pairs where a significant difference is found. The combinations with no significant difference in influence on pupil dilation can be used for designing test environments, since this will not affect pupil dilation as much. These combinations can be found in Table 3.5. In terms of pupil dilation on colors, $\boldsymbol{H_{04}}$ **can be rejected and thus** $\boldsymbol{H_{14}}$ **is true**.

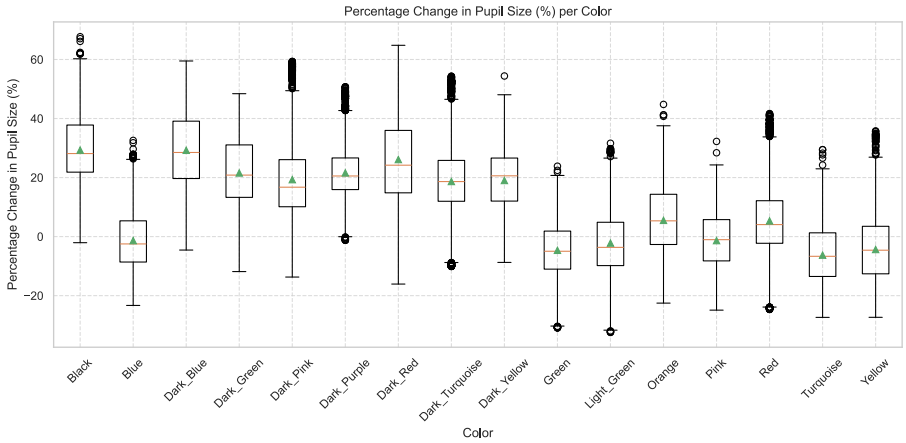


Figure 3.7: The percentage of pupil dilation by color. This is the percentage of dilation of the pupil's size in contrast to the median pupil size captured during the white wall stare calibration.

A (R,G,B)	B (R,G,B)	Am	Bm	W	Z	p	r
Black (0,0,0)	Dark Green (3,107,5)	25.52	19.99	9.0	-2.73	$p < 0.01$	0.80
Black (0,0,0)	Dark Pink (145,3,112)	25.52	16.14	13.0	-2.48	$p < 0.05$	0.72
Black (0,0,0)	Dark Purple (97,3,112)	25.52	20.10	13.0	-2.48	$p < 0.05$	0.65
Black (0,0,0)	Dark Turquoise (8,107,107)	25.52	17.61	6.0	-2.92	$p < 0.01$	0.56
Black (0,0,0)	Dark Yellow (139,128,0)	25.52	19.56	10.0	-2.67	$p < 0.001$	0.65
Dark Blue (3,5,107)	Dark Green (3,107,5)	28.06	19.99	4.0	-3.05	$p < 0.01$	0.84
Dark Blue (3,5,107)	Dark Pink (145,3,112)	28.06	16.14	10.0	-2.67	$p < 0.01$	0.65
Dark Blue (3,5,107)	Dark Purple (97,3,112)	28.06	20.10	14.0	-2.42	$p < 0.05$	0.62
Dark Blue (3,5,107)	Dark Turquoise (8,107,107)	28.06	17.61	3.0	-3.11	$p < 0.01$	0.62
Dark Blue (3,5,107)	Dark Yellow (139,128,0)	28.06	19.56	0.0	-3.30	$p < 0.01$	0.78
Dark Purple (97,3,112)	Dark Turquoise (8,107,107)	20.10	17.61	18.0	-2.17	$p < 0.05$	0.77
Dark Red (107,5,5)	Dark Turquoise (8,107,107)	23.41	17.61	14.0	-2.42	$p < 0.05$	0.75

Table 3.3: Significance values between dark colors regarding influence on pupil dilation. Only combinations with significant differences are mentioned. Am is the median of A and Bm is the median of B.

A (R,G,B)	B (R,G,B)	Am	Bm	W	Z	p	r
Blue (73,171,245)	Red (255,0,0)	-1.42	3.41	13.0	-2.48	$p < 0.05$	0.52
Blue (73,171,245)	Turquoise (64,237,237)	-1.42	-8.36	8.0	-2.80	$p < 0.01$	0.64
Green (0,255,0)	Orange (255,128,0)	-4.77	4.47	6.0	-2.92	$p < 0.01$	0.49
Green (0,255,0)	Red (255,0,0)	-4.77	3.41	7.0	-2.86	$p < 0.01$	0.56
Light Green (166,255,0)	Orange (255,128,0)	-3.80	4.47	6.0	-2.92	$p < 0.01$	0.73
Light Green (166,255,0)	Red (255,0,0)	-3.80	3.41	15.0	-2.35	$p < 0.05$	0.42
Orange (255,128,0)	Pink (232,130,237)	4.47	-1.74	11.0	-2.61	$p < 0.01$	0.62
Orange (255,128,0)	Turquoise (64,237,237)	4.47	-8.36	1.0	-3.23	$p < 0.01$	0.68
Pink (232,130,237)	Red (255,0,0)	-1.74	3.41	7.0	-2.86	$p < 0.01$	0.71
Pink (232,130,237)	Turquoise (64,237,237)	-1.74	-8.36	8.0	-2.79	$p < 0.01$	0.85
Red (255,0,0)	Turquoise (64,237,237)	3.14	-8.36	1.0	-3.23	$p < 0.01$	0.70
Red (255,0,0)	Yellow (255,235,4)	3.14	-2.26	14.0	-2.42	$p < 0.05$	0.52

Table 3.4: Significance values between light colors regarding influence on pupil dilation. Only combinations with significant differences are mentioned. Am is the median of A and Bm is the median of B.

A (R,G,B)	B (R,G,B)
Light	
Blue (73,171,245)	Orange (255,128,0)
Blue (73,171,245)	Green (0,255,0)
Blue (73,171,245)	Light Green (166,255,0)
Blue (73,171,245)	Pink (232,130,237)
Blue (73,171,245)	Yellow (255,235,4)
Green (0,255,0)	Light Green (166,255,0)
Green (0,255,0)	Pink (232,130,237)
Green (0,255,0)	Turquoise (64,237,237)
Green (0,255,0)	Yellow (255,235,4)
Light Green (166,255,0)	Pink (232,130,237)
Light Green (166,255,0)	Turquoise (64,237,237)
Light Green (166,255,0)	Yellow (255,235,4)
Orange (255,128,0)	Red (255,0,0)
Orange (255,128,0)	Yellow (255,235,4)
Pink (232,130,237)	Yellow (255,235,4)
Turquoise (64,237,237)	Yellow (255,235,4)
Dark	
Black (0,0,0)	Dark Blue (3,5,107)
Black (0,0,0)	Dark Red (107,5,5)
Dark Blue (3,5,107)	Dark Red (107,5,5)
Dark Green (3,107,5)	Dark Pink (145,3,112)
Dark Green (3,107,5)	Dark Purple (97,3,112)
Dark Green (3,107,5)	Dark Red (107,5,5)
Dark Green (3,107,5)	Dark Turquoise (8,107,107)
Dark Green (3,107,5)	Dark Yellow (139,128,0)
Dark Pink (145,3,112)	Dark Purple (97,3,112)
Dark Pink (145,3,112)	Dark Red (107,5,5)
Dark Pink (145,3,112)	Dark Turquoise (8,107,107)
Dark Pink (145,3,112)	Dark Yellow (139,128,0)
Dark Purple (97,3,112)	Dark Red (107,5,5)
Dark Purple (97,3,112)	Dark Yellow (139,128,0)
Dark Red (107,5,5)	Dark Yellow (139,128,0)
Dark Turquoise (8,107,107)	Dark Yellow (139,128,0)

Table 3.5: The colors that were not found to be significantly different in terms of influence on the pupil dilation.

3.6 Conclusion Of Pupil Influences

When designing a virtual environment for assembly task training, depth does not have any effect on pupil dilation according to the depth test. When the colors involved have no significant differences in illumination, pupils will not specifically react to the depth of an object.

The used colors, on the other hand, can have a significant influence. As seen in the color test, some colors result in significant differences in pupil dilation. This means that if a training environment has a lot of colors with significant differences, the pupil dilation tends to react more, masking the pupil dilation for cognitive load.

Regarding brightness, it is good to know how cognitive load can be measured with more variable lighting conditions during the training for the brightness test. However, adding this variation in scene brightness would seem unnecessary when training employees and trying to find the process's difficulties. It is, however, handy to use the luminance calculations so you always have an idea of what luminance the trainee's eyes may be exposed to. Virtual depth is found not to influence pupil dilation. This is verified since the pupils no longer responded to the virtual depth of the block after the difference in luminance was eliminated, meaning the pupils just respond to the colors and the lightness of the environment. Regarding the colors, there are colors found to have no significant difference in effect on pupil dilation, while other color combinations do have different effects. In designing the task, we must ensure that colors with the same influence are used wherever possible.

Chapter 4

Measuring Cognitive Load In Virtual Assembly Environments Using Eye Tracking Data

A new user study is set up after exploring the effects on pupil dilation and blinking rates. This user study focuses on measuring cognitive workload in a more complex virtual task, which lies closer to a task performed in a real job. This chapter explains what user study was performed and to which results it led.

The task performed during the study is placing stickers on an assembled electrical cabinet, which is a task within an existing software tool created by UHasselt EDM. This is done in a training phase and then again in the assembly phase. The electrical cabinet is an excellent assembly to use in this study because it is complex but yet simple enough for participants to understand its structure. It also requires careful attention since some components are exactly the same but still need a different label on them, thus ensuring participants do not have it too easy. The electrical cabinet consists of three rails, as seen in Figure 4.1. Each of these rails was considered to be built in the user study. The third rail, which is at the back, was not chosen because it is repetitive. The repetitiveness would come from the rail needing two objects to be placed next to each other six times. This might make participants feel under-aroused and bored. The second rail is also not an option, since it consists of just four items to be placed. This means the study would be too short to extract valuable data. Lastly, rail one was a great option. It has some variety, even though the gray terminal must be placed many times. It was also not a very short task, meaning valuable data could be extracted. The stickers were chosen above rail one because placing them might be repetitive, just like the previous tasks, but the stickers contain text, meaning participants cannot simply take a swift look at the image on the instructions sign and match the image. Participants are hereby forced to read the sticker's content, thus introducing cognitive workload.

4.1 Study Setup

4.1.1 Material

Participants were given a Varjo XR-3 mixed-reality headset to conduct the test. As mentioned in the previous chapter, this headset has excellent hand-tracking capabilities and can measure eye metrics such as pupil dilation, gaze direction, and blinks. Besides the measurements, participants were also asked to fill in a NASA-TLX after both tests, which allows for finding correlations between the measured data and the perceived workload scores. The participant's age, eye color, gender, and whether they wear glasses were also collected. The study is split into

two different parts. Both parts had 12 participants, thus a total of 24 people participated in the study. The two groups are needed in order to compare the cognitive workload measurements between the two different levels of assistance.

4.1.2 Participants Demographics

Figure 4.1 shows the demographics of the participants for this study. Note, "have glasses" indicates the participant generally wears glasses, while "required glasses" means the participant had to wear them during the study in order to properly see. All participants in this study were students in Computer Science at Hasselt University, of which two were PhD students.

Group	participant	males	females	have glasses	average age	required glasses
High-assistance	12	10	2	6	24.58	1
Low-assistance	12	11	1	6	23.33	2

Table 4.1: Table displaying the participant demographics. For clarification, "have glasses" means they generally wear glasses, and "required glasses" means they had to wear them during the test in order to see.

4.1.3 The Phases

The user study consists of three phases: the practice phase, the training phase, and the assembly phase. In the **practice phase**, participants have full assistance and are tasked with assembling rail one. The goal, however, is not to have the participants assemble the rail, but rather to try out different actions inside the VR environment, like scrolling through the resource list, spawning an item, placing an item, and more, to get accustomed to the workings. In the **training phase**, participants complete the sticker task with full assistance enabled. This phase is the same for both the **High-assistance** and the **Low-assistance** group. Finally, the **assembly phase** is different for the two groups. Here, the **High-assistance** group completes the sticker test again with **full assistance**, while the **Low-assistance** group completes it with a lowered form of assistance.

For the **High-assistance group**, the assistance level remains the same to study if there is a measurable cognitive load difference between their training phase and their assembly phase, because during training, participants are told they have to repeat the process, meaning they have to remember what they have to do. In contrast, during the assembly phase, participants figure that they get the same assistance and manage to finish the assembly more swiftly. This should result in the assembly phase having lower cognitive load levels than the training phase.

The **Low-assistance** group has a lower level of assistance to study whether the cognitive load goes up during the assembly phase, since they must recall where the sticker goes. Besides, the two different groups can be compared to see if there is a significant difference in cognitive workload between the group with full assistance both times and the group with lower assistance in the assembly phase.

4.1.4 Difference Between Assistance Levels

With **Full assistance**, the instructions interface displays the step that needs to be performed, as seen in Figure 4.2. Besides, the resource interface will display a green border around the resource that needs to be spawned, as seen in Figure 4.3. Also, the hologram displayed on the target location shows exactly what sticker needs to be placed, plus the rotation it needs to be in. When a participant is holding the wrong resource, a red outline will also appear around the object, indicating that it is not the correct object. Conversely, a green outline suggests that the participant holds the proper object. The outline can be seen in Figure 4.4.

For the **lower assistance**, the hologram on the target location is a circle that does not display the shape and rotation of the sticker. Also, the resource highlight in the resource interface is disabled, so users have to search for themselves. Besides, the correct or incorrect resource highlights around the object are disabled, ensuring participants do not get help knowing whether they have the wrong sticker. Lastly, the instruction interface does not display any information, meaning participants can no longer quickly look at the picture of the sticker and match it. The participants can only see the target hologram and must remember which sticker is supposed to go at that location. This is designed to induce as much load a possible.

4.1.5 The Virtual Environment

For the user study, virtual assembly software for electric cabinets made in Unity was used. This software was created internally by Hasselt University. This software was then adapted to be used with the Varjo headset and to comply more with the results of the pupil dilation exploration. Some changes were made to the skybox, which was changed to a plain gray wall to ensure the background had only one color. This gray is also the same shade of gray that most of the items of the electric cabinet have. The colored highlights were also changed. The pupil exploration found that the green shade, which indicated the correct object was being held, and the red shade, which indicated the wrong object was being held, had significant differences in pupil dilation during the color test. Therefore, these colors were changed to the darker versions used in the color test, which were found not to have a significant difference in effect on pupil dilation. Unnecessary user interface elements, such as the UI for selecting the assembly task and the assistance level, were removed. This is done so that participants cannot accidentally select assemblies or assistance levels that are not part of the test.

The electric cabinet was put on a baseplate. This plate can be seen in Figure 4.1. On top of the leftmost object, a green hologram can be seen. In full assistance mode, this indicates where the stickers should be placed. Behind the plate, there is a screen displaying the steps. This screen can be seen on Figure 4.2. Next to the instructions UI, there is the resource UI. This UI displays all the resources required to assemble the electric cabinet. Participants can scroll through this list to find the needed resources and spawn them by snapping their index and thumb together. Figure 4.3 shows an image of this UI. Lastly, a participant may spawn in the wrong object. Therefore, a trash bin was added, allowing the participants to throw the object away so that it no longer floats around in the virtual environment. This bin can be seen on Figure 4.4.

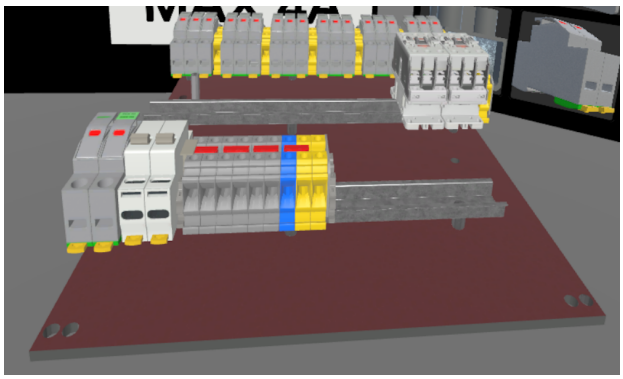


Figure 4.1: The electrical cabinet on which participants have to place stickers. On the first row, left, a green highlight can be seen. This indicates where the sticker should be placed for full assistance.

Figure 4.2: The instructions interface. This interface displays a short title of the task, a task description, and an image of the resource that needs to be placed on the electric cabinet.

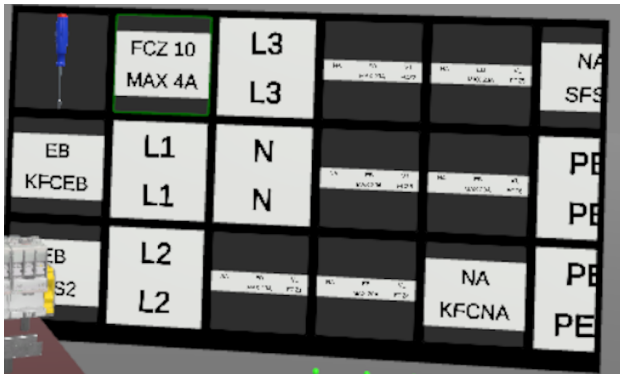


Figure 4.3: The resource interface. This interface displays all resources that are required to assemble the electric cabinet. Participants can scroll through the list and spawn the objects they need.

4.2 Calibration And Baseline Process

Eye calibration is required before starting the training and assembly phases. This calibration process consists of three steps. First, the eye measurements are calibrated using Varjo's built-in 5-point calibration process. This short calibration step consists of participants looking at a point in different locations without moving their heads. This allows the headset to follow the eyes and measure its metrics more precisely. After this, the size of the eyes for a luminance of zero is measured by having participants stare at a black wall for ten seconds from a certain location. This is done because it is required for the brightness correction calculations, which will be needed in case pupil dilation seems to be affected too much by the black of the instruction and resource interfaces, since these contrasts pretty hard with the environment. The black wall is the instructions interface without any content. Finally, the general baseline is calculated by having participants stare at the built assembly for ten seconds from the same location. This baseline is used to calculate the Z-score of pupil dilation change throughout the test, thus showing if pupils shrank or expanded during the test compared to the baseline.

4.3 Task Design

As previously mentioned, the tasks for both the training and the assembly phase consist of placing stickers in the right location of the electric cabinet. 27 stickers need to be placed on the electric cabinet; thus, the task consists of 27 steps. There are a total of 19 distinct stickers that

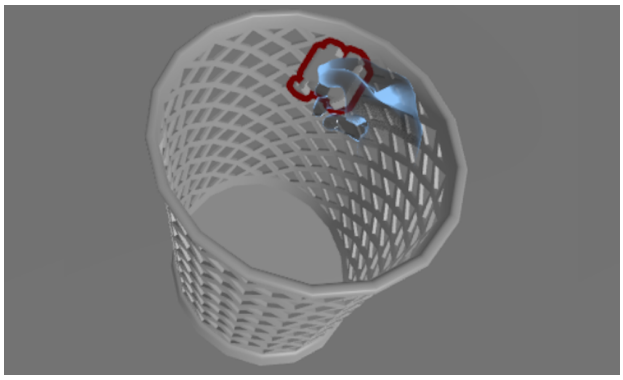


Figure 4.4: The trash bin. When a participant spawns in the wrong object. The object can be put in this bin, which will destroy it.

need to be used. For every step, a green target hologram is displayed at the location where it should be placed. Besides, depending on the assistance level, the instructions for every step are also displayed in the instructions interface. The participant's task is to view the instructions (when applicable) and figure out which sticker needs to be placed down, find the correct sticker in the resource interface, and finally, spawn it in and place it on the electric cabinet to proceed to the next step. The stickers that need to be placed can be found in Table 4.2.

Rail 1		Rail 2		Rail 3	
Step	Sticker	Step	Sticker	Step	Sticker
1	FCZ 10 MAX 4A	13	NA KFCNA	16	PEN PEN
2	NA SFS1	14	EB KFCEB	17	PEN PEN
3	EB SFS2	15	PE PE	18	PEN PEN
4	L1		/	19	PEN PEN
5	L1		/	20	PEN PEN
6	L2		/	21	PEN PEN
7	L2		/	22	NA EB V1 MAX 20A FCZ1
8	L3		/	23	NA EB V1 MAX 20A FCZ2
9	N N		/	24	NA EB V1 MAX 20A FCZ3
10	PEN PEN		/	25	NA EB V1 MAX 20A FCZ4
11	PE PEN		/	26	NA EB V1 MAX 20A FCZ5
12	XDK1		/	27	NA EB V1 MAX 20A FCZ6

Table 4.2: Table showing all stickers that need to be placed on each of the three rails

4.4 Measurements

In the study, multiple measurements are being recorded. These measurements are either measured by the VR headset, the training tool, or measured through a questionnaire. The subjective measurements are recorded in the NASA-TLX questionnaire. Here, mental demand, physical demand, temporal demand, performance, effort, and frustration are verbally questioned and noted. Besides this, the VR headset also captures multiple metrics such as pupil dilation, blinks, gaze direction, the object the participant is looking at, the RGB values of said object at the location of the raycast's hit, and the calculated luminance score. Pupil dilation is a primary measurement used to study whether pupil dilation changes in cognitive workload changes in the assembly tasks. Pupil dilation will be measured and transferred into Z-scores. To calculate these Z-scores, the pupil dilation for both the training and assembly uses the data from the training phase calibration. This is done to base both data on the same base data, where the

participant has no knowledge of the test. Otherwise, the second calibration might introduce biased data. Besides, papers state that the blinking rate can also indicate cognitive workload, which is why blinks are identified in the data based on eye-tracking status. A blink is identified when the eye tracking of both eyes is visible, but the two data points before it are not visible. The gaze direction is also measured but not necessarily used, since the object the participant is looking at is recorded. Knowing exactly which object the participant is looking at can help better understand what the participant is doing. For example, it can show how many times and for how long a participant was looking at the instructions. Finally, the RGB values and calculated luminance of where the participant is looking are recorded as well. Besides all these metrics, the training tool also reports when the test starts and ends, every new step, every resource spawned, whenever a resource is binned, and every finished step. This allows for measuring the time each step took, the number of wrong objects spawned, and the total duration of the phase.

4.5 Hypothesis

In this study, several hypotheses will be tested. The within-group hypotheses are checked separately for each group, while the group comparison hypotheses are used to compare the two groups.

4.5.1 Within Group

Since different literature states that the **pupil dilation** increases when cognitive load is induced [EHR21; PS03], the hypothesis in terms of pupil dilation is the following:

- H_{05} The pupil dilation is **not** significantly **different** during the assembly phase compared to the training phase
- H_{15} The pupil dilation is significantly **different** during the assembly phase compared to the training phase

Next, the **blinking rate** is recorded to see if the cognitive workload affects the blinking rate when the assistance level remains the same. Here, it is unclear whether the blinking rate should increase or decrease. Therefore, the following hypotheses are made:

- H_{06} There is **no** significant difference in blinking rate between the training phase and the assembly phase
- H_{16} There is a significant difference in blinking rate between the training phase and the assembly phase

4.5.2 Group Comparison

Both groups should experience the **assembly phase** differently, **pupil dilation** should be influenced by this experience difference. Since the low-assistance group needs to think more, more cognitive load should be generated, resulting in the following hypothesis:

- H_{07} The pupil dilation is **not** significantly **bigger** during the assembly phase of the low-assistance group compared to the assembly phase of the high-assistance group
- H_{17} The pupil dilation is significantly **bigger** during the assembly phase of the low-assistance group compared to the assembly phase of the high-assistance group

For the **blinking rate**, the question is whether there is a difference in blinks per minute between the two groups or not. This is needed to declare whether the blinking rate is an indication of cognitive load or not. The following hypothesis is created for the blinking rate:

- H_{08} There is **no** significant difference in the blinking rate of the high-assistance group's assembly phase and the assembly phase for the low-assistance group

- H_{18} There is a significant difference in the blinking rate of the high-assistance group's assembly phase and the assembly phase for the low-assistance group

4.6 Study Procedure

Before the start of the study, a pilot test was conducted. This pilot study aims to determine whether everything is set up correctly and if the study is feasible in a reasonable amount of time. For this study, the pilot study consisted of two participants. One participant had full assistance in the assembly phase, but the other had no assistance. This meant that no instructions were shown, no highlights were shown, and not even the holograms were shown. The participant got frustrated and could not finish the task because it was far too complex to do with just one training session. This then led to low-assistance being used further in the study.

After the pilot study was finished, the actual study could begin. Participants were gathered by walking around the building and asking people to participate in the user study. Upon entering the room, participants were greeted, seated, and shown a form of informed consent stating what the study is about, a short description of the task, what data is gathered, and what happens with their data. After giving consent, the participant was asked their age and whether they wore glasses. Meanwhile, their eye color was also being noted. Next, before bringing the participants into the VR environment, participants were given a paper document explaining how to use the VR tool in picture form. This document visually displays how to scroll in the resource interface, spawn an item, and place an item. Verbal explanation was also given to the participants to make it even more straightforward and ensure they fully understood how it works.

After the data gathering and the explanation, participants were given the headset and placed in the VR environment to start the practice phase. In this phase, participants are asked to scroll through the resource list, spawn an item, turn an item around, throw away an item, and place down an item. Here, the participants are also told that when letting an item loose will not drop the item. The calibration steps were also present in the practice phase. It is not used in any data measurements, but it is still performed so the participant knows what to do when the calibration does matter. The participant could try out different actions as long as required and was instructed to mention when they were ready to proceed to the training phase.

After the participant had approved it, the training phase started. Participants were told they had to remember what they did because the assembly would be repeated, and then went through the calibration steps. After, they waited three seconds for a red button to pop up, which they needed to press to start the test. Now, the participants learn step-by-step which stickers need to be placed at which locations. This is done with the highest form of assistance to aid them in the learning process. After each sticker was placed and the assembly was completed, participants were asked each question of the NASA-TLX, which they would then verbally answer.

After the NASA-TLX questionnaire for the training phase was filled in, the assembly phase began (which was different for both groups). The participants were put back into the virtual environment and were tasked with placing the stickers again. Just like in the training phase, the participants go through the calibration procedure and must press the red button, after which the task starts. Finally, just like in the training phase, participants were asked to answer each question of the NASA-TLX verbally.

After this questionnaire was completed, participants were asked to hand back the VR headset and were thanked for participating in the user study.

Chapter 5

Results

This chapter discusses the results of the user study conducted to test the predefined hypotheses. To test whether the data was normally distributed or not, a **Shapiro test** is used. The **Wilcoxon test** is used to verify whether there are significant differences between the **training phase and assembly phase data** within each group. Finally, the **Mann-Whitney U** test is used to check for significant differences between the two groups.

5.1 Hypothesis 5: Pupil Dilation

Hypothesis 5 consists of H_{05} and H_{15}, which question whether **pupil dilation** is significantly **different** during the assembly phase compared to the training phase. All measured pupil sizes are converted to a Z-score of change compared to the median captured value during the calibration step. This is done for both the training phase and the assembly phase. However, the measured pupil sizes for the assembly phases are not compared to their own calibration, but to the calibration of the training phase. This is done to base both phases on the exact same data, which makes a better comparison than trying to compare values with different bases.

5.1.1 High-Assistance

Figure 5.1 shows the change in pupil dilation in Z-score for each step in both phases. Blue boxes indicate the values for the training phase, and red boxes indicate values for the assembly phase.

Since pupil dilation should decrease with less cognitive load, the data is first checked to see if the assembly phase's median pupil dilation is smaller compared to the median of the training phase. Table 5.3 shows the difference between different measured metrics for the training and assembly phases in the high-assistance group. Green numbers indicate that the value is following the expected trend in the assembly phase compared to the training phase. Controversially, red numbers indicate that the value is the opposite of the expected trend. The first column of this table displays the difference between the median pupil change, which is in Z-score, of the full test. This is calculated by doing the median pupil dilation change for assembly (in Z-score) minus the median pupil dilation change for training (in Z-score). This means it is not a Z-score itself, but the difference between the two Z-scores. The expected trend is that participants will notice the second time is also with the same high-level assistance as in the training phase, thus resulting in smaller pupils. As seen in the table, 9 of the 12 participants (75%) have an actual decrease in pupil size in the assembly phase compared to the training phase. Besides, 3 of the 12 participants have an increase in pupil dilation. When taking the overall median value between all participants, the value results in -0.315, which means that overall, the pupil sizes have **decreased** in the **assembly phase**, as expected.

Wilcoxon is run for each step separately and for the test overall. The task consists of a total of 27 steps. Five of these steps have a significant difference in pupil dilation change between training and assembly, found with a Wilcoxon signed-rank test. These steps are shown in Table 5.1.

Steps with significant differences between phase						
Step	training median	assembly median	W	Z	p	r
L3 L3_1	-2.62	-4.08	13.0	-2.04	$p < 0.05$	0.59
PE PE_1	-2.53	-2.37	10.0	-2.28	$p < 0.05$	0.66
PEN PEN_2	-2.51	-3.34	11.0	-2.20	$p < 0.05$	0.63
PEN PEN_5	-5.00	-5.50	9.0	-2.67	$p < 0.01$	0.77
NA EB V1 MAX 20A FZC1_1	-1.54	-2.91	5.0	-2.67	$p < 0.01$	0.77

Table 5.1: Table displaying the steps that were found to have a significant difference in completion time for the **high-assistance** group.

On the overall data, the medians of the training phase and assembly phase were -2.08 and -2.83, respectively. The Wilcoxon signed-rank test showed **no significant difference in pupil dilation change** between the training and assembly phase ($W = 19, Z = -1.57, p > 0.05, r = 0.45$). This means that even though the pupil generally decreased for most participants, it is not a significant difference. This hints that the only real effect on the cognitive load here is the learning effect. For the rest, there is no difference in influence op cognitive workload. This means that H_{05} **cannot be rejected** for the **high-assistance group**.

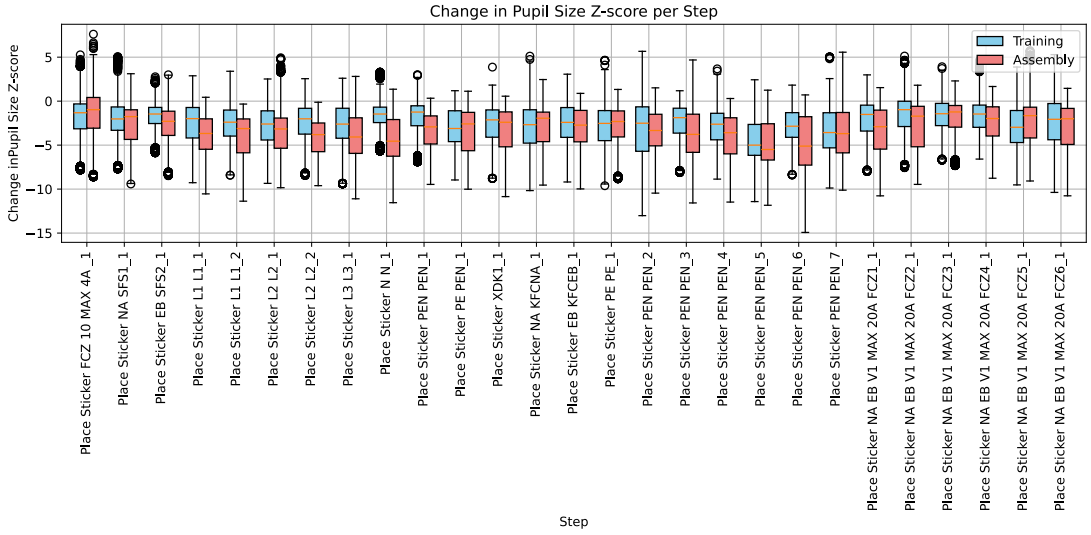


Figure 5.1: A boxplot showing the values of pupil dilation change in Z-score across all high-assistance participants grouped per task for both phases. Training phase in blue and assembly phase in red.

5.1.2 Low-Assistance

Figure 5.2 shows the change in pupil dilation in Z-score per step for both phases. Blue boxes are data from the training phase, and red boxes represent assembly phase data.

First, it is important to check whether the pupil dilation actually went up in the assembly phase compared to the training phase, since pupil sizes should increase with more cognitive load. The difference in median pupil dilation between the two phases can be seen in the bottom half of Table 5.3. As seen in the table, ten of the twelve participants did indeed have an increase

in pupil dilation. The remaining two participants had a decrease in pupil dilation, of which only one had a greater difference than 1. This means that, in general, the pupil dilation is indeed bigger in the assembly phase compared to the training phase. This makes sense since participants did not have a lot of assistance in this round and thus were required to access their memory.

Similar to the high-assistance group, the Wilcoxon test is first done for each step separately, and then overall. The task for this group is the same as for the other group, thus also having 27 steps. Unlike the high-assistance group, with five of 27 steps having significant differences in pupil dilation, the low-assistance group has twelve significant steps, which are shown in Table 5.2.

Steps with significant differences between phase						
Step	training median	assembly median	W	Z	p	r
L1 L1_1	-1.48	-1.61	1.0	-2.98	$p < 0.001$	0.86
L1 L1_2	-1.70	-2.80	9.0	-2.35	$p < 0.05$	0.68
L2 L2_1	-1.17	-1.13	10.0	-2.28	$p < 0.05$	0.66
L3 L3_1	-1.46	-1.04	13.0	-2.04	$p < 0.05$	0.59
NA EB V1 MAX 20A FCZ1_1	-0.99	-0.84	9.0	-2.35	$p < 0.05$	0.68
NA KFCNA_1	-1.38	-0.86	10.0	-2.28	$p < 0.05$	0.66
PE PE_1	-1.26	-0.63	8.0	-2.43	$p < 0.05$	0.70
PEN PEN_2	-0.98	-0.64	9.0	-2.35	$p < 0.05$	0.68
PEN PEN_3	-1.41	-0.60	3.0	-2.82	$p < 0.01$	0.82
PEN PEN_4	-1.24	-1.12	12.0	-2.12	$p < 0.05$	0.61
PEN PEN_6	-1.52	-1.09	10.0	-2.28	$p < 0.05$	0.66
XDK1_1	-1.44	-0.46	7.0	-2.51	$p < 0.01$	0.73

Table 5.2: Table displaying the steps that were found to have a significant difference in completion time for the **low-assistance** group.

One thing that stands out here is that L3 L3_1, PE PE_1, PEN PEN_2, and NA EB V1 MAX 20A FZ1_1 have significant differences in both groups. However, PEN PEN_5 seems to have lost its significant differences in this group. This means that the significance in the other group might not have to do with this step being difficult, but rather having other causes.

The medians of the training and assembly phases were -1.27 and -1.05, respectively. The Wilcoxon signed-rank test showed a significant difference in pupil dilation between the different tasks of ($W = 9, Z = -2.35, p < 0.05, r = 0.68$). Also, as seen in Table 5.3, P25 is the only participant with a smaller pupil dilation in the assembly phase with a difference bigger than one, which might hint at an outlier. Outliers were identified to be P15, P23, and P25, in which P15 and P23 were positive outliers, while P25 was a negative outlier. If the outliers were to be excluded, the Wilcoxon signed-rank test shows that the effect in phase would increase to ($W = 0.0, Z = -3.06, p < 0.01, r = 0.88$). Besides, if only the negative outlier P25 were excluded, the Wilcoxon signed-rank test shows a significant effect of ($W = 0.0, Z = -3.06, p < 0.001, r = 0.88$). However, even though there is an outlier pulling the results towards less significance, there is still a significant difference in pupil dilation between the two phases. This means that H_{05} **can be rejected** for the **low-assistance group**, meaning H_{15} is the true hypothesis.

5.2 Hypothesis 6: Blinking Rate

Hypothesis 6 consists of H_{06} and H_{16}, which question whether there is a significant difference in **blinking rate** or not between the two phases. Before checking for significance, it is important to know whether the blinking rate is smaller or higher in the assembly phase compared to the training phase. For now, this is not really known because Schwerd *et al.* states that the blinking rate goes down when cognitive load goes up [SS24], while Biondi *et al.* states that it goes up when cognitive load goes up [Bio+23].

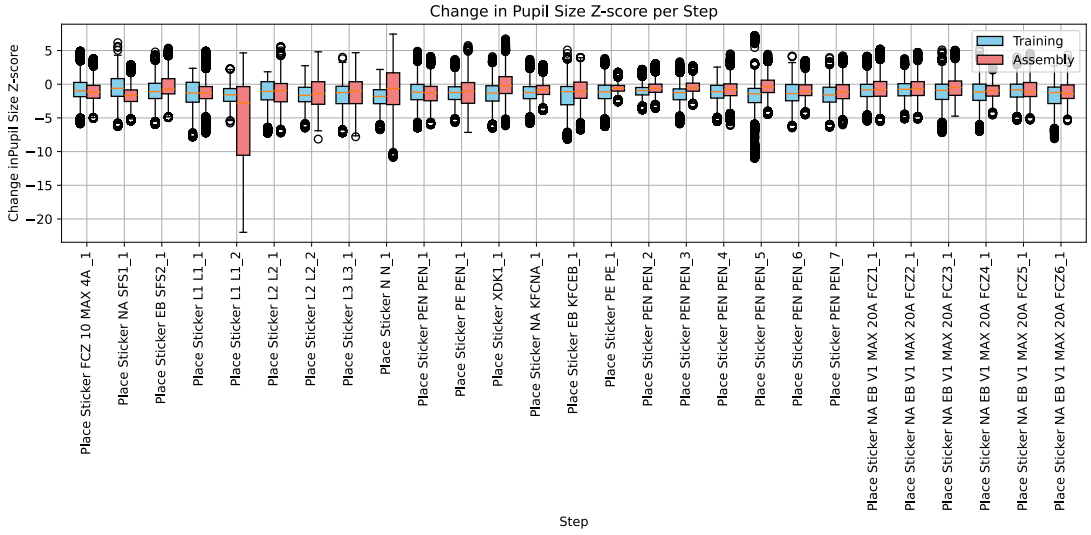


Figure 5.2: A boxplot showing the values of pupil dilation change in Z-score across all low-assistance participants grouped per task for both phases. Training phase in blue and assembly phase in red.

5.2.1 High-Assistance

Figure 5.3 displays the median of blinks per minute from the training and assembly phases for each participant in the **high-assistance group**. Here, the assembly value is lower seven times, higher three times, and the same height twice. These differences between the phases are also clearly displayed in Table 5.3 in the second column, which from this point has yellow numbers to indicate a difference in either direction with a value less than one. The figure and table indicate that in most cases (58%) the participant blinked fewer times. However, a noticeable point is the massive spike in the assembly test of P5. As seen on Table 2, it is clear that P5 started blinking very much in the 3rd and the last two minutes. This is possibly because P5 has spent a lot of time in VR, since P5's training time was 14 minutes. It is possible that P5's eyes were getting fatigued from the VR headset after spending so much time in it.

The median blinks per minute for the training and assembly phase were 16.62 and 14.89, respectively. The Wilcoxon signed-rank test revealed that there is **no significant difference in blinking rate** between the two phases with values ($W = 38, Z = -0.08, p > 0.05, r = 0.02$). This means that the null hypothesis H_{06} **cannot be rejected** for the **high-assistance group**.

5.2.2 Low-Assistance

Figure 5.4 shows the median of blinks per minute for both the training and assembly phases for all participants of the **low-assistance group**. The difference between these values can be seen in the second column of the bottom part of Table 5.3. Compared to the **high-assistance group**, where **seven of the twelve participants** had a **lower blinking rate** in the assembly phase, the **low-assistance group had nine** participants who **increased** their blinking rate in the assembly phase. This means that the blinking rate might actually increase with cognitive load, since both **pupil dilation and blinking rate increased** in the assembly phase. This must be confirmed in a correlation test that is discussed later in this thesis.

On the blinking rate data for the training and assembly phase in the **low-assistance group**, the medians were 16.5 and 20.5, respectively. The Wilcoxon signed-rank test showed a significant effect of phase ($W = 12, Z = -2.12, p < 0.05, r = 0.61$). This means that the increase in blinking rate in the assembly phase is actually significantly different from the training phase.

This shows that not only did pupil dilation increase, but also the blinks per minute. Since the blinking rate significantly differs, hypothesis H_{06} **can be rejected** for the **low-assistance group**.

5.3 Gerenal Participant Behaviour

Next to the head measurements of **pupil dilation and blinking rate**, there are some other behavioural measurements that can give an indication of cognitive load change. These metrics are, however, small indications and cannot be used to actually indicate cognitive load itself, but rather indicate that a phase seems to be harder or easier.

5.3.1 Completion Time

An indicator like this can be the **completion time of the phase**. Figure 5.5 displays the completion time for the training and assembly phase for all participants of the **high-assistance group**. The difference between these completion times is displayed in the third column of Table 5.3. As seen in both the figure and the table, all participants completed the assembly phase in fewer seconds compared to the training phase, which is the expected trend. On average, the assembly phase is completed 43.45% faster compared to the training phase. This is three minutes and twenty-one seconds. It is also noticeable that only two out of 12 participants did not finish at least one minute faster. As seen in Table 5.3, the smallest difference in completion time greater than one minute is a difference of one and a half minutes, which is still quite large.

The median test durations for the training and assembly phases are 445.92 seconds and 230.02 seconds, respectively. The Wilcoxon signed-rank test shows a significant effect on phase ($W = 0.0, Z = -3.06, p < 0.001, r = 0.88$), meaning the assembly phase was finished **significantly faster** compared to the training phase for **high-assistance group participants**.

Figure 5.6 shows the completion time for the training and assembly phase for all participants of the **low-assistance group**. The differences between these values can also be seen in Table 5.3. The figure shows that 10 of the 12 participants took longer to complete the assembly phase compared to the training phase. Surprisingly, two participants managed to complete the assembly phase faster. One of these two participants even completed the assembly phase 115 seconds faster. On average, the assembly phase took $69,60\%$ longer to complete. This is 5 minutes and 20 seconds. This indicates that the completion time follows the expected trend of being longer in the assembly phase. Figure 5.7 displays the median number of seconds it took for each step to be completed for both the training and the assembly phase. Here, the assembly phase shows massive bottlenecks in some of the steps, meaning participants struggled there. This figure can help display where bottlenecks are in the assembly process and where further training is required.

On the **low-assistance** assembly times, the median phase durations for the training and assembly phases were 415.35 and 589.89 seconds, respectively. The Wilcoxon signed-rank test revealed a significant effect of ($W = 7.0, Z = -2.51, p < 0.01, r = 0.72$), which indicates that the completion time for the assembly phase took **significantly longer to complete** compared to the training phase.

5.3.2 Time Spent Looking At Instructions

Another one of these indicators is the **time in seconds that participants spent looking at the instructions**, which only exists for the **high-assistance group**. Figure 5.8 shows the number of seconds each participant in the **high-assistance group** has viewed the interface instructions in both the training and assembly phases. The differences between the values for the training and assembly phases can be seen in the fourth column of Table 5.3. The figure and table do not provide data for P2. This is because P2 was part of a pilot study in which the

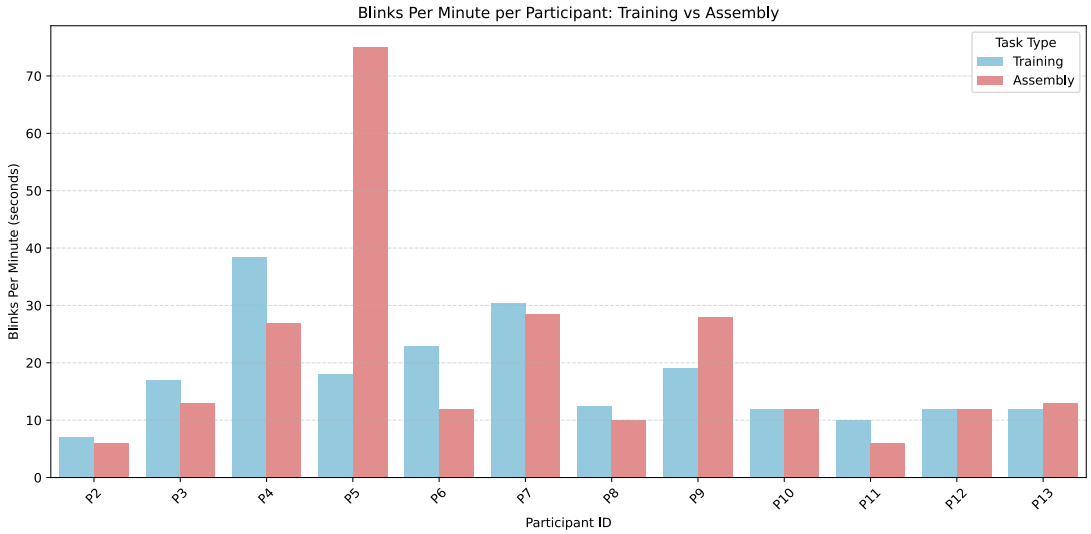


Figure 5.3: A bar plot showing the median of blinks per minute across all **high-assistance** participants for both phases. Training phase in blue and assembly phase in red.

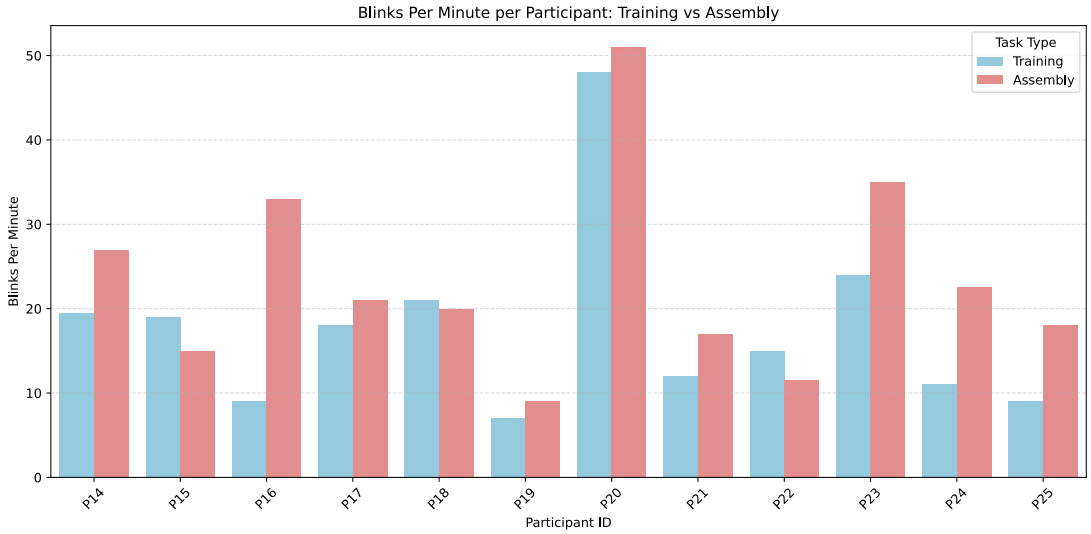


Figure 5.4: A bar plot showing the median of blinks per minute across all **low-assistance** participants for both phases. Training phase in blue and assembly phase in red.

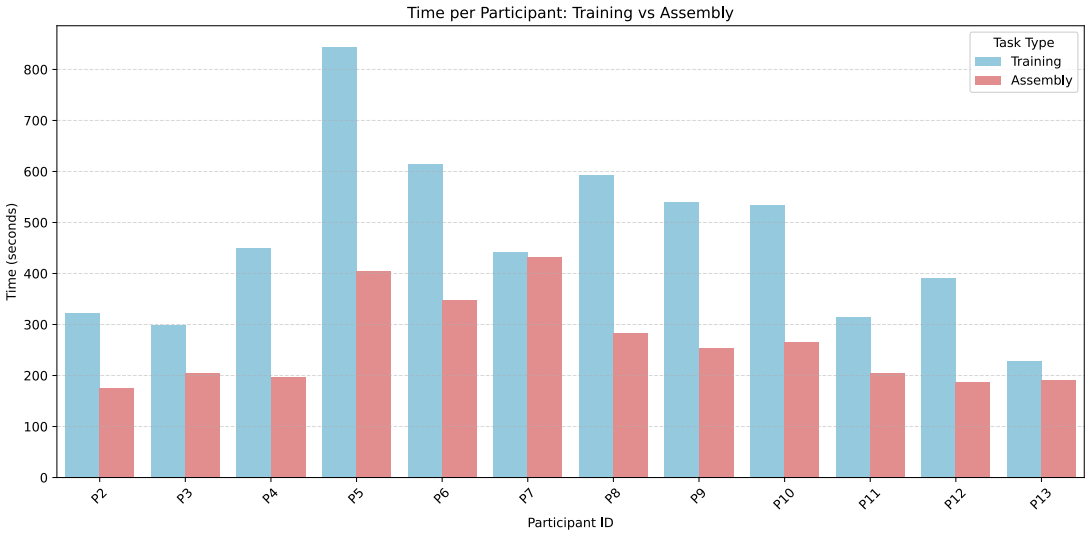


Figure 5.5: A bar plot showing the completion time in seconds across all **high-assistance** participants for both phases. The training phase is in blue, and the assembly phase is in red.

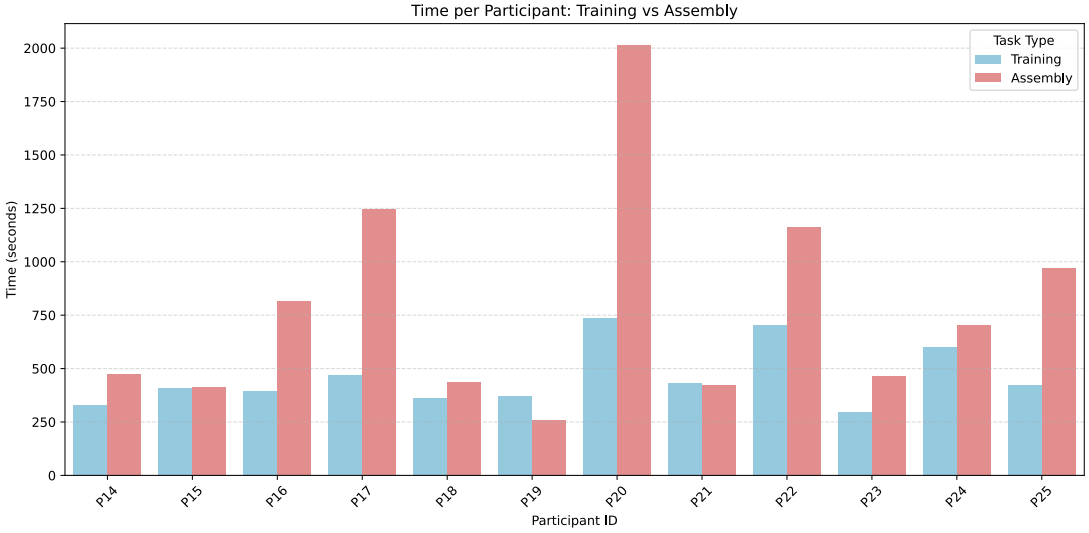


Figure 5.6: A bar plot showing the completion time in seconds across all **low-assistance** participants for both phases. Training phase in blue and assembly phase in red.

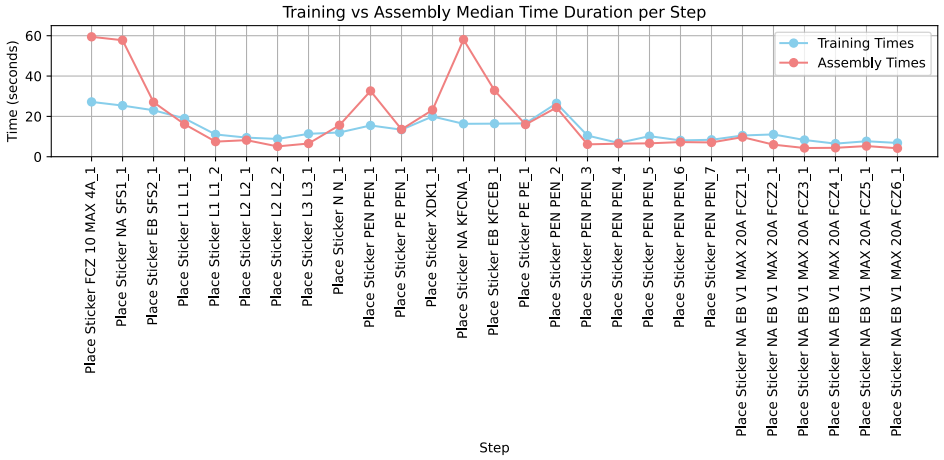


Figure 5.7: A plot showing the completion time in seconds of every step across all low-assistance participants for both phases. The training phase is in blue, and the assembly phase is in red.

feature was not yet implemented. In the figure and table, it is clear that all participants spent less time looking at the instructions in the assembly phase, compared to the training phase, which is as expected. In fact, participants spent 59.12% fewer seconds viewing instructions on average. This is around 22 seconds. It is also noticeable that P9, who completed the assembly phase more than four and a half minutes faster, spent one minute and seven seconds less looking at instructions. This shows that participants make less use of the provided instructions because they know what to do and just require a quick look at the sticker's image to know the next step.

On the seconds spent looking at instructions, the median time in seconds for the training and assembly phase was 27.47 and 14.36, respectively. The Wilcoxon signed-rank test has revealed a significant effect of phase ($W = 0.0, Z = -2.93, p = 0.001, r = 0.8847$), indicating that participants spent **significantly less time looking at instructions** in the assembly phase compared to the training phase. This could also indicate that the assembly phase seemed easier to the participants than the training phase.

5.3.3 Time Spent Looking At Resource interface

The final indicator is the **time in seconds spent looking at the resource interface**. Figure 5.9 shows the number of seconds each participant spent looking at the resource interface. The difference between these values can be seen in the fifth column of Table 5.3. Here, P2 also has no data for P2 since this was not yet implemented. This data shows that all participants spent fewer seconds looking for resources in the assembly phase compared to the training phase. On average, participants spent $34,18\%$ seconds less, which is 54 seconds. This indicates that participants find the assembly phase easier.

On the seconds spent looking at the resource interface for the **high-assistance group**, median values of the training and assembly phase are 159.03 and 101.46, respectively. A Wilcoxon signed-rank test showed a significant effect of phase ($W = 0.0, Z = -2.93, p = 0.001, r = 0.88$), meaning that the **time spent searching for the right resources** has **significantly decreased** in the assembly phase, which also hints that participants could find the right resources more easily.

Figure 5.10 shows the number of seconds each participant viewed the resource interface in the **low-assistance group**. The difference between these values is displayed in the fifth column of Table 5.3. This table shows that, in contrast to the high-assistance group, where every

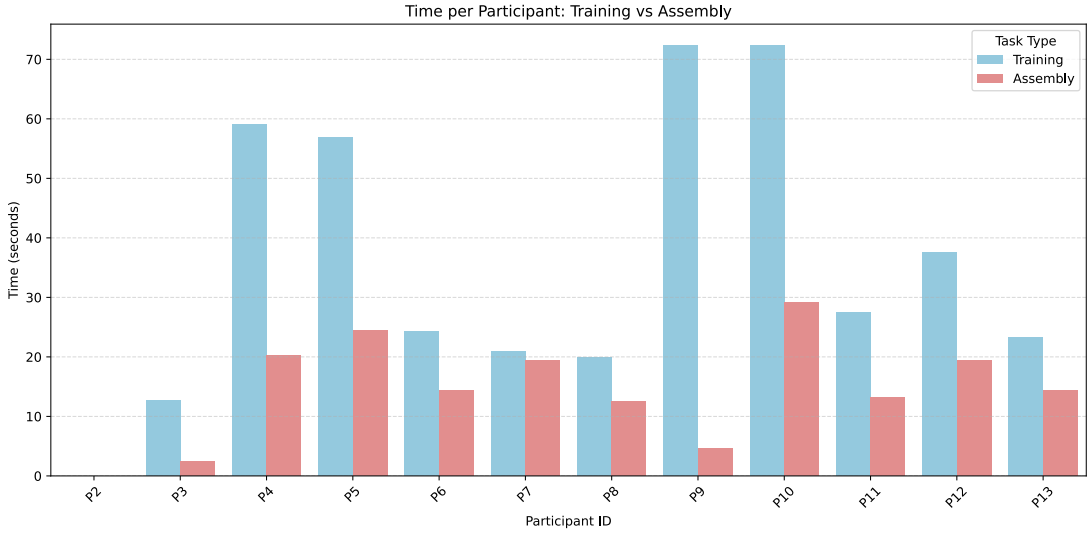


Figure 5.8: A bar plot showing the time spent looking at the instructions panel in seconds across all high-assistance participants for both phases. Training phase in blue and assembly phase in red. Note that P2 has no data since this was not implemented yet in the pilot study.

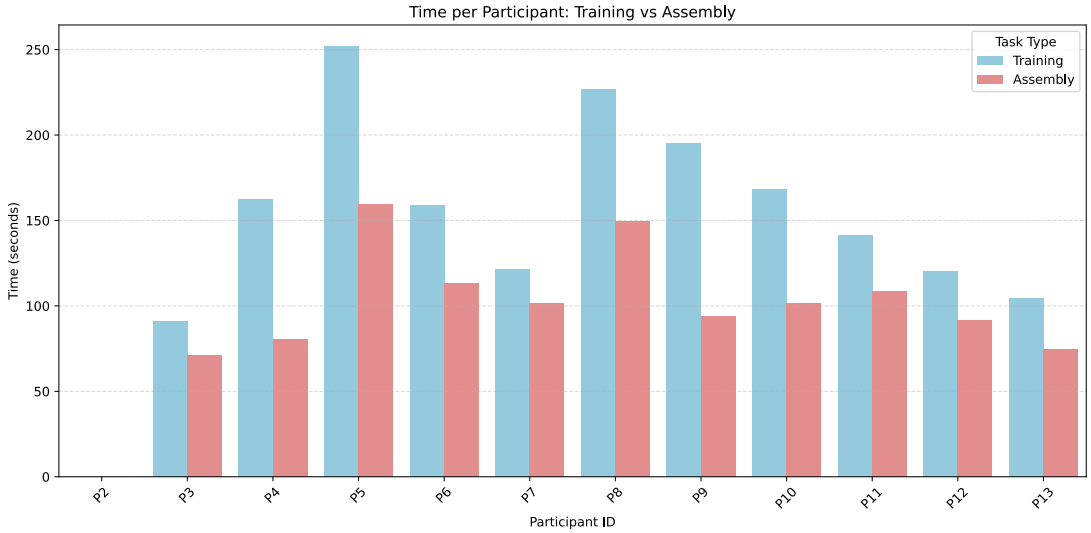


Figure 5.9: A bar plot showing the time spent searching for the correct resources in seconds across all high-assistance participants for both phases. The training phase is in blue, and the assembly phase is in red. Note that P2 has no data since this was not implemented yet in the pilot study.

participant spent less time looking at the resource interface, all low-assistance group participants spent more time looking for the right resources. On average, low-assistance participants spent $95,92\%$ more seconds looking at the resource interface. This means that the number of seconds almost doubled. This indicates that participants found it more difficult to find the correct resources in the assembly phase.

For the time spent looking at the resource interface in the **low-assistance group**, the median values for training and assembly phase are 161.40 and 228.83, respectively. A Wilcoxon signed-rank test reveals a significant effect in phase ($W = 0.0, Z = -3.06, p < 0.001, r = 0.88$), which means that the **low-assistance group** spent a **significantly longer time** looking for the right resources in the assembly phase compared to the number of seconds searching for the resources in the training phase.

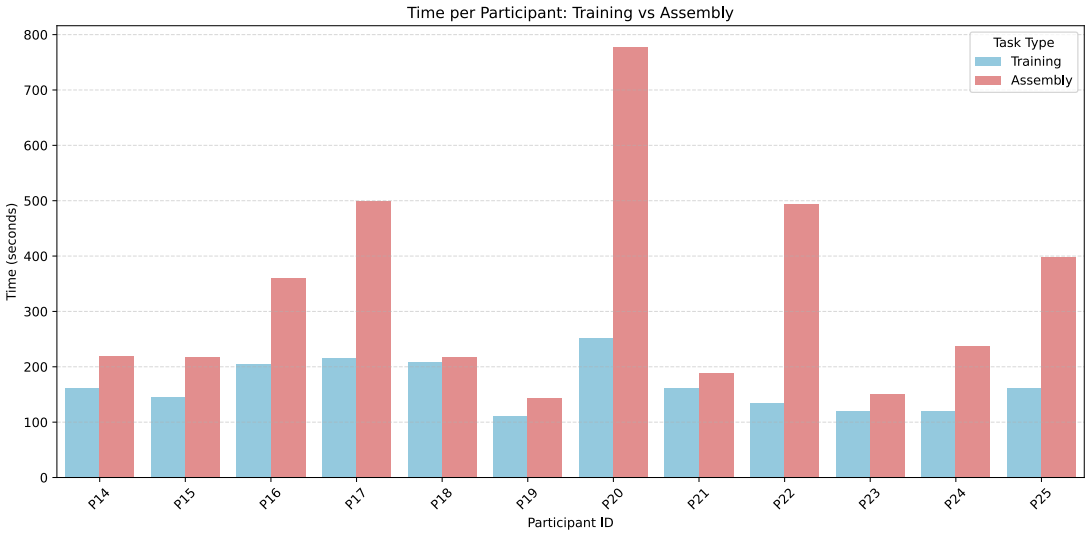


Figure 5.10: A bar plot showing the time spent searching for the correct resources in seconds across all low-assistance participants for both phases. Training phase in blue and assembly phase in red.

Participant	**Difference of metrics between training and assembly phase**				
	Dilation (Z-score diff)	Blinks (pm)	Duration (s)	Instructions (s)	Resource (s)
High-assistance					
P2	-0.387	-1	-145.967	/	/
P3	-0.942	-4	-94.297	-10.090	-19.767
P4	0.429	-11.5	-253.255	-38.866	-81.994
P5	-0.047	57	-438.910	-32.383	-92.385
P6	-0.599	-11	-265.516	-9.853	-45.858
P7	-3.974	-2	-9.370	-1.364	-19.919
P8	-0.312	-2.5	-309.570	-7.315	-77.222
P9	-0.317	9	-285.406	-67.569	-101.236
P10	-0.819	0	-268.887	-43.138	-66.887
P11	0.455	-4	-109.694	-14.304	-32.888
P12	0.417	0	-203.497	-18.245	-32.769
P13	-0.183	1	-37.386	-8.968	-28.497
Low-assistance					
P14	0.505	7.5	145.408	/	58.121
P15	1.763	-4	4.744	/	73.318
P16	0.385	24	419.106	/	155.858
P17	0.944	3	778.004	/	283.633
P18	-0.042	-1	74.919	/	10.200
P19	0.516	2	-115.430	/	31.129
P20	0.025	3	1276.807	/	526.012
P21	0.230	5	-9.635	/	27.585
P22	0.614	-3.5	457.699	/	360.344
P23	2.629	11	167.858	/	31.449
P24	0.453	11.5	102.781	/	117.862
P25	-1.043	9	546.288	/	236.730

Table 5.3: This table displays the difference between the values for the training and the assembly test. Dilation (Z-score) indicates the difference between the median change in dilation for the training and the assembly test. Blinks (pm) is the difference between the blinks per minute of the training test vs the assembly test. Duration (s) is the difference between the durations of the two tests in seconds. Instructions (s) are the difference in the number of seconds viewed at the instructions interface. Lastly, Resource (s) is the difference in the number of seconds viewed at the resource interface. Green values indicate that the assembly's value was following the expected trend, red means the assembly's value was the opposite of the expected trend, and yellow indicates a difference in any direction smaller than 1 (Does not apply for dilation Z-score).

5.4 Correlations

Finally, after finding which of the metrics significantly differ between the two phases, it is also interesting to see if any of the metrics interact with each other. This interaction is tested by using a Pearson correlation test on all of the different metrics. A Pearson correlation (or Pearson's r) measures the linear relationship between two variables. It is used to indicate how strongly and in what direction two variables are related to each other. The range for this test is from -1 to 1, where +1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 means no relationship. For this study, many correlations are tested for both phases. Important tested correlations are the following:

- Each NASA-TLX metric
  - Pupil dilation
  - Blinking rate
  - Test duration
  - Time spent looking at instructions
  - Time spent searching for resources
- Test duration
  - Pupil dilation
  - Blinking rate
  - Time spent looking at instructions
  - Time spent searching for resources
- Blinking rate
  - Pupil dilation
  - Time spent looking at instructions
  - Time spent searching for resources
- Time spent looking at instructions
  - Pupil dilation
- Time spent searching for resources
  - Pupil dilation

5.4.1 High-Assistance

For the high assistance group, only a very small number of checked correlations are found to actually be correlated. These few correlated metrics and their correlation trend are displayed on Figure 5.11. The metrics shown are R, P-value, Power, and BF10. R is the correlation value in the range of -1 to 1, P-value indicates whether the correlation is significant or not, Power indicates the power of the given correlation in which a number closer to 1 indicates that the result is more trustable, and BF10 is the Bayer Factor, which indicates which hypothesis is most likely true. For the Bayes Factor, a value over 1 indicates that the alternative hypothesis is most likely true, and a value of one or lower indicates no evidence to reject the null hypothesis. Figure 5.11 shows all the found correlations.

The first correlation, at the top left op Figure 5.11 shows a positive relation between **the number of mistakes made and blinking rate** found in the **assembly phase**. A mistake is identified when a participant spawns in the wrong resource. The correlation tells us that participants who made **more mistakes** seemed to have a **higher blinking rate**.

The next significant correlation is a correlation between the **physical demand score given by participants and the pupil dilation** in the **training phase**, which is important in terms of cognitive workload measurement. This can be seen at the top right of 5.11 This negative correlation suggests that participants who rated **more physical demand** had a **lower pupil dilation**, while participants who rated a **lower physical demand** had **higher pupil dilation**. The median pupil dilation Z-score for both phases can also be seen in Table 5.4.

The next significant correlation, shown at the middle left of 5.11, is between **The frustration score** given by participants in the NASA-TLX and the **Blinking Rate** in the **training phase**. This positive correlation states that participants who rated higher frustration blinked more. This correlation is also significant within the first 7 minutes of the **training phase**. This would suggest that frustration would make participants blink more often.

Another correlation shown in Figure 5.11 is between the **frustration score and duration of the test** in the **assembly phase**. This positive correlation states that whenever a participant took longer to complete the assembly phase, the test frustration score is higher, which contradicts the idea that participants simply take their time and finish it at a lower pace with less cognitive load. It suggests that participants mostly took longer to finish their **assembly phase** because something frustrating happened, possibly the hand tracking working annoyingly, or the sticker not snapping instantly.

The final significant correlation on 5.11 is a positive correlation between **test duration and blinking rate** in the **assembly phase**. This correlation states that participants who required more time to finish the assembly phase generally blinked more. A possible reason for this could be that the participants' eyes got more fatigued from using the VR headset. Figure 5.3 displays the median number of blinks for each participant in the **high-assistance group** for both phases, and Table 5.4 shows these values in text form.

Since it is known which correlations are relevant, it is important to view two correlations that did not turn out to be significant, namely, pupil dilation change, together with looking at instructions and searching for resources. The reason why it is crucial for these to be not significant is that these interfaces mostly consist of a black background, which is not a common color in the assembly setup. Therefore, it will influence the pupil dilation. This correlation **not** being significant indicates that participants looking at the interfaces for a longer time did **not** significantly affect their pupil dilation. However, pupil dilation also does not really seem to significantly correlate with any of the measured metrics except for **physical demand**. Yet, it might be interesting to take a look at some of the pupil dilation correlations to see if something interesting can be found.

First, **pupil dilation and mental load score** is checked, since mental load should be affecting pupil dilation. In the **training phase**, the data is very scattered. Some participants who gave high mental load scores have high pupil dilation, but others have also rated higher mental load while having some of the lowest pupil dilation of all participants. Besides, two participants with mental load scores 3 and 4 both have a pupil dilation change value of above -1, indicating the pupil dilation was pretty high. Two participants who rated mental load a 9 and 13 have roughly the same high pupil dilation. A participant with a mental load score of 2 also has roughly the same pupil dilation as a participant who rated 15, and the two participants with the most pupil shrinkage both rated 12 on the mental load score. This means that, unfortunately, the **training phase data is not giving any valuable insight**. This extends to the **assembly phase** data as well. The participant with the highest mental load score has the second highest pupil dilation, while the others in the top 5 rated 5, 3, 2, and 2, respectively to their ranking.

Next, it might also be interesting to look at the correlation between **pupil dilation and physical demand** in the **assembly phase**, since its training phase had a significant correlation. Even though the assembly phase is the exact same for this group, the assembly phase correlation is a lot more variable. In the training phase, a clear pattern is shown, but the assembly phase lacks this pattern completely. This results in the assembly phase data not really giving any value, unlike the training phase data.

Also, **pupil dilation and effort score** are reviewed to see if there is any indication that participants who felt like they had to put in more effort actually had more pupil dilation. In the **training phase**, the data points are, like the previous reviews, enormously scattered and cannot give any indication of the true relation between pupil dilation and the effort the participant felt they had to put in. The participant with the highest effort score of 18 has a higher than average pupil dilation, but the participant with the second highest score of 16 has the second most pupil **shrinkage** of all. All the other, except for one, have way higher pupil dilation and lower scores, so the **training phase** data fails to give any indication here. When looking at the **assembly phase data**, it also does not improve the insights. Although the two participants with the highest effort score are in the top half of pupil dilation rankings, so are four participants who rated 2, 2, 3, and 3. Besides, the participant with the highest rating also has smaller pupil Z-scores compared to the participant with the second highest score. The participant with the second highest score has roughly the same dilation Z-score as the four participants with very low effort ratings. Since participants with roughly the same eye measurements have extremely different opinions on effort, it **cannot provide any insight**.

Finally, **pupil dilation and frustration score** is checked to see if there could be any valuable insight. The **training phase** correlates slightly downwards, which indicates that more frustration would lead to smaller pupils. However, the correlation cannot be found to be significant evidence because the data points are rather scattered. The participant with the highest frustration score of 18 has the second most pupil contraction, but the participant with a score of 17 has the 5th highest dilation. Besides, the top 5 highest pupil dilation values gave a frustration rating of 6, 8, 8, 2, 17, respectively, according to their ranking. Therefore, it is only a slight indication because the **training phase** data alone cannot give any useful information. The **assembly phase** shows a steeper negative correlation. This is, however, clearly because of an outlier, where a participant with the lowest pupil dilation has the only rating of 20. The rest of the scores are all smaller or equal to 13 and while the lowest pupil dilation is smaller than -7, all the rest are at least bigger than -6. The other points are still pretty scattered and do **not give any meaningful information** about whether pupil dilation actually shrinks because of frustrating encounters.

After reviewing these, it seems like pupil dilation fails to give any good indication of explaining its behavior, which is not ideal when measuring cognitive load through pupil dilation. Perhaps the **low-assistance** group will give better insight since their assembly phase differs from the training phase.

5.4.2 Low-Assistance

Same as for the high-assistance group, it is interesting to see whether there are correlations between the different metrics. The interaction is tested through a Pearson correlation test. The same metrics are tested as were done for the high-assistance group, except for the ones involving the instruction interface, because the assembly phase does not show instructions on it. The significant correlations can be seen in Figure 5.12.

The first significant correlation, at the top left corner of Figure 5.12, is a positive correlation between **mistakes made and the duration of the test** in the **training phase**. This correlation indicates that participants who made more mistakes in their training phase also required a longer time to finish the assembly. This is because, when they are not aware they made a mistake, they try to place the wrong sticker for some time.

The next correlation, which is found to be significant, can be seen at the top right corner of Figure 5.12. This is also a positive correlation between **mistakes and duration of the test** in the **assembly phase**. This correlation is thus present in both phases for the **low-assistance group**. However, the correlation is a lot stronger in the **assembly phase**, which is because participants generally made a lot more mistakes in this phase due to the lack of instructions, making them spend more time on more wrong stickers.

Next, there is a positive correlation between **mistakes and blinking rate** in the **training phase**. This correlation also exists in the **high-assistance group**, but then it is for their **assembly phase**. The correlation states that participants who made more mistakes generally blinked more. It is, however, important to note that this correlation seems to be dependent on an outlier who has a relatively high blinking rate compared to the rest of the participants.

The next correlation found in the right middle of Figure 5.12 is a positive correlation between **Effort score and mistakes** in the **assembly phase**. This correlation shows that participants who made more mistakes felt like more effort was required to finish the jobs. This is because they mostly could not remember the order of the stickers, which caused them to have to put in more effort to finish the task.

The next correlation, found at the bottom left of Figure 5.12, is a negative correlation between **performance score and percentage of time spent looking at the resource interface** in the **training phase**. This correlation shows that participants who spent more time looking at the resource list felt like they were more aware of the task and performed better.

The final correlation is a negative correlation between the **duration of the test and the percentage of time spent looking at the resource interface** in the **training phase**. This correlation states that participants who took longer to complete the training spent less time looking at the resource list.

There are no significant correlations found with **pupil dilation**, which indicates that pupil dilation did not go down or rise with any of the other measured metrics. This is not ideal since this is the base metric. However, the pupil dilation correlations with **mental load, physical demand, effort, and frustration** will be looked into to check if there is any indication of correlations between the scores and pupil dilation.

First, **pupil dilation and mental load** is checked. In the **training phase**, it seems like participants with relatively close Z-scores range their mental load scores from one all the way to 10. The correlation goes slightly downwards, but only because of two participants whose Z-scores are lower than 4. Without these, the data points roughly seem to make a hill-like shape, with more pupil dilation indicating more mental demand up to score 5, and then it goes down again. This makes it hard to get any indication of the relation between the two metrics. In the **assembly phase**, it looks like there is a very rough indication that would say participants with lower pupil dilation rated more mental load. However, there is a wide confidence interval ($CI = [-0.76, 0.38]$), meaning the direction and strength are unsure.

The next non-significant correlation to review is **pupil dilation and physical demand**. The high-assistance group training phase correlation for this was significant, but the **low-assistance** group is less ideal to gather information from. In the **training phase**, the general trend would be pretty flat if it weren't for an outlier with really high pupil shrinkage (r = 0.05). This means that, unlike the high-assistance group, the low-assistance group did not even give a slight indication of reproducing their significant correlation. For the **assembly phase**, the same applies as for the training phase. The correlation is extremely flat when the outlying point is not considered. This means that pupil dilation did **not relate** to physical demand for this group.

Next, there is the non-significant correlation between **pupil dilation and effort score**. In the **training phase**, it looks like a very slight positive trend is shown because of an outlier, which is the same participant as for the others. Without this participant's data, the trend goes flat. Participants with similar pupil dilation Z-scores also have different scores ranging from two all the way to twelve. This means that the pupil dilation and effort clearly do not relate in the training phase. In the **assembly phase**, however, there is a very slight indication of a negative effect. It is not true for the entirety of the data points, but six of the points sort of show a negative trend, which would mean participants who felt like they put in more effort would have smaller dilation.

Finally, there is the non-significant correlation between **pupil dilation and frustration score**.

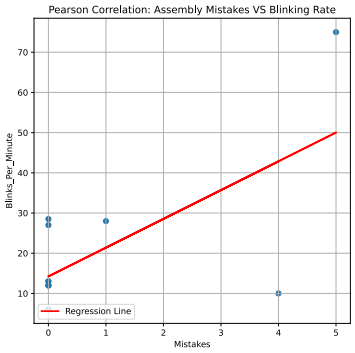
In the **training phase**, there generally seems to be two hills. Pupil dilation change rose for the participant who rated frustration a 5, compared to the one who rated it a 2. Yet, pupil dilation went down again for the participants of ratings 7 and 8, and 11. After this, participants who rated a 12 generally have higher pupil dilation again, after which another peak happens at a score of 17. Then, the participant with a frustration score of 20 has lower pupil dilation again. This means that pupil dilation did not seem to really relate to the frustration participants felt. The **assembly phase** data is also not very informative, with a flat line and a lot of points not showing any order. This means that the frustration also cannot explain much about the pupil's behavior.

Participant	**Dilation**		**Blink Rate**		**Duration**		**Resource**	
	Train	**Assembly**	**Train**	**Assembly**	**Train**	**Assembly**	**Train**	**Assembly**
P2	-0.957	-1.134	7	6	322	176	/	/
P3	-3.072	-4.013	17	13	300	206	91	71
P4	-1.537	-1.109	38	27	450	197	162	80
P5	-5.854	-5.901	18	75	844	405	252	159
P6	-1.733	-2.332	23	12	614	348	159	113
P7	-3.343	-7.316	30.5	28.5	442	433	121	102
P8	-0.871	-1.184	12.5	10	593	283	227	149
P9	-0.686	-1.003	19	28	540	255	195	94
P10	-0.595	-1.414	12	12	535	266	168	102
P11	-2.932	-2.477	10	6	315	206	141	109
P12	-5.258	-4.841	12	12	391	188	120	92
P13	-4.080	-4.263	12	13	229	191	105	75

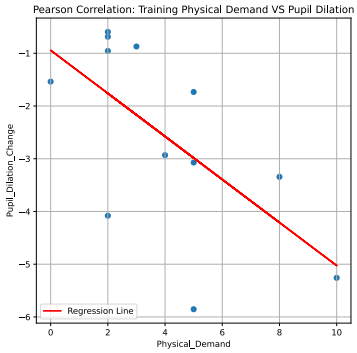
Table 5.4: Table displaying the median values for both training and assembly phase for the different metrics. This data is from the **high-assistance** group. Dilation is in Z-score, blinking rate is in blinks per minute, duration is in seconds, and resource is in seconds (the correlations use percentage as $(ResourceSeconds - Duration) * 100$).

Participant	**Dilation**		**Blink Rate**		**Duration**		**Resource**	
	Train	**Assembly**	**Train**	**Assembly**	**Train**	**Assembly**	**Train**	**Assembly**
P14	-11.705	-11.199	19.5	27	329	474	162	220
P15	1.148	2.911	19	15	408	412	145	218
P16	-0.934	-0.539	9	33	395	814	204	360
P17	-2.217	-1.273	18	21	468	1246	215	499
P18	-2.479	-2.521	21	20	361	436	208	218
P19	-0.839	-0.322	7	9	373	258	112	143
P20	-0.793	-0.767	48	51	737	2014	251	777
P21	-2.614	-2.383	12	17	433	423	161	189
P22	-1.485	-0.870	15	11	705	1163	134	495
P23	-5.513	-2.884	24	35	295	462	120	151
P24	0.939	1.392	11	22.5	603	705	120	238
P25	-1.565	-2.609	9	18	423	969	162	399

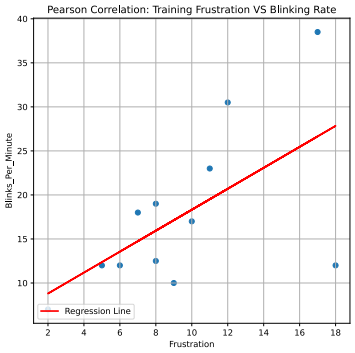
Table 5.5: Table displaying the median values for both training and assembly phase for the different metrics. This data is from the **low-assistance** group. Dilation is in Z-score, blinking rate is in blinks per minute, duration is in seconds, and resource is in seconds (the correlations use percentage as $(ResourceSeconds - Duration) * 100$).



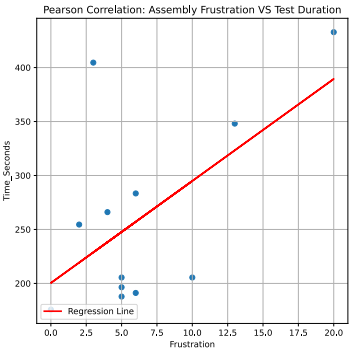
R: 0.657183
CI95%: [0.13, 0.89]
p-val: 0.020226
BF10: 3.974
Power: 0.685755



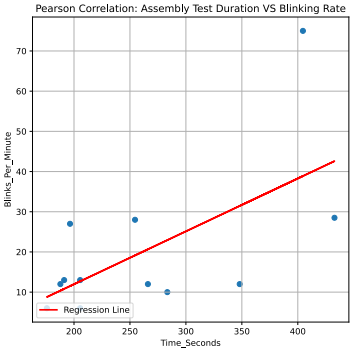
R: -0.634359
CI95%: [-0.89, -0.1]
p-val: 0.025462
BF10: 3.328
Power: 0.649657



R: 0.602106
CI95%: [0.04, 0.87]
p-val: 0.0383
BF10: 2.438
Power: 0.580117

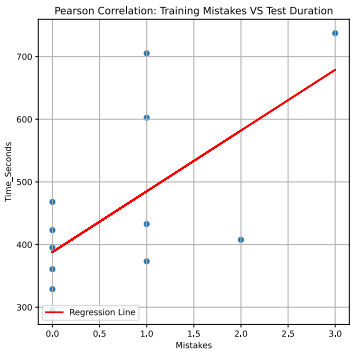


R: 0.580681
CI95%: [0.01, 0.87]
p-val: 0.047732
BF10: 2.066
Power: 0.539997

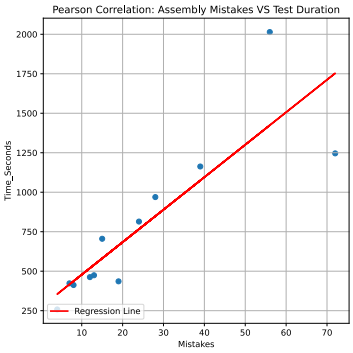


R: 0.610802
CI95%: [0.06, 0.88]
p-val: 0.034881
BF10: 2.618
Power: 0.596632

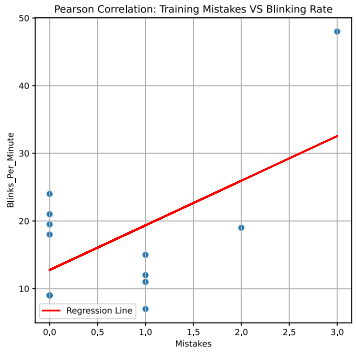
Figure 5.11: Graph displaying a visual plot of the interaction between different measured metrics for the **high-assistance** group. The red line is the trend line. The only graphs considered significant Pearson Correlations.



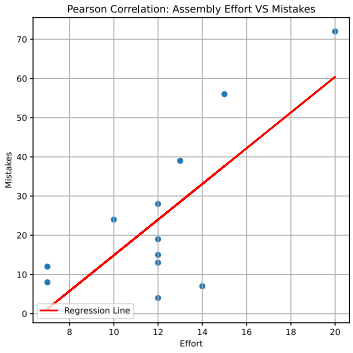
R: 0.649803
CI95%: [0.12, 0.89]
p-val: 0.022187
BF10: 3.7
Power: 0.671553



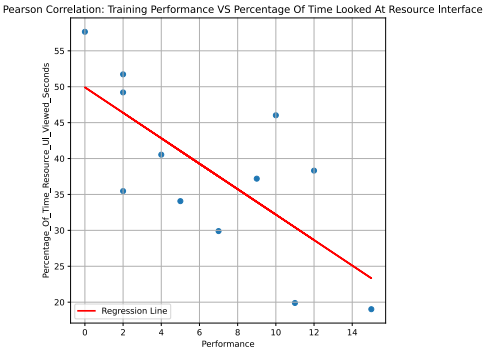
R: 0.861759
CI95%: [0.57, 0.96]
p-val: 0.000314
BF10: 108.751
Power: 0.979751



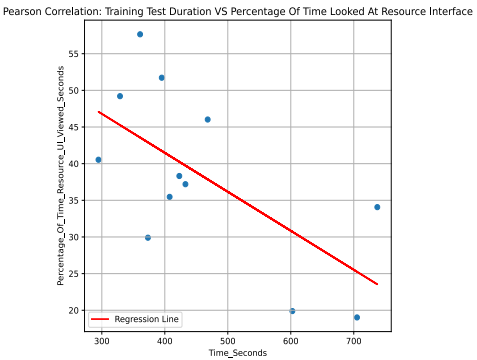
R: 0.580449
CI95%: [0.01, 0.87]
p-val: 0.047842
BF10: 2.062
Power: 0.539568



R: 0.74803
CI95%: [0.31, 0.92]
p-val: 0.005145
BF10: 11.612
Power: 0.850393



R: -0.71675
CI95%: [-0.91, -0.24]
p-val: 0.008715
BF10: 7.661
Power: 0.79708



R: -0.64622
CI95%: [-0.89, -0.11]
p-val: 0.023182
BF10: 3.577
Power: 0.664651

Figure 5.12: Graph displaying a visual plot of the interaction between different measured metrics for the low-assistance group. The only graphs considered significant Pearson Correlations.

5.5 Group Comparison

Finally, to truly analyze the difference in cognitive load between the two groups, the pupil dilation, blinking rate, and test duration for the training and assembly phases are compared across the two groups. The results of this are discussed in this section.

5.5.1 Hypothesis 7: Pupil Dilation

Hypothesis 7 consists of H_{07} and H_{17}, which question whether pupil dilation in the assembly phase is significantly **bigger** for the low-assistance group compared to the high-assistance group or not.

To compare both groups, both the median pupil dilation Z-score of the assembly phase are compared. A Mann-Whitney U test is used to see whether there are significant differences in pupil dilation between the two groups. The Mann-Whitney U test revealed a significant difference in pupil dilation ($U = 5.52, p < 0.001, RBC = -4.47, CLES = 0.27$). Here, the RBC value shows that **high-assistance** members generally have **lower** pupil dilation compared to **low-assistance** members. Besides, CLES shows that, when taking a random sample from group the high-assistance and low-assistance group, there is a 27% chance that the high-assistance member has higher pupil dilation.

Figure 5.13 shows the median Z-scores in a boxplot. The figure shows that the **low-assistance group** their values start and reach higher compared to the dilation values of the **high-assistance group**, who had the same assistance level twice and mostly had a shrinkage in pupil size in the assembly phase compared to the training phase. This finding means that hypothesis H_{07} **can be rejected**.

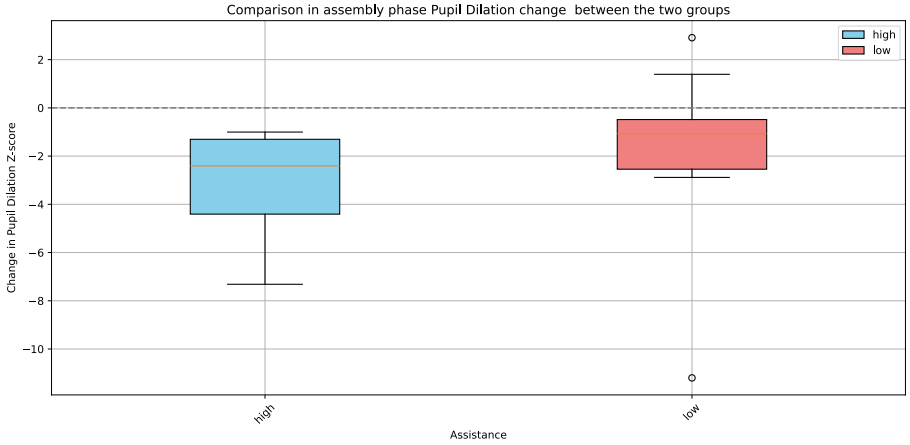


Figure 5.13: Boxplot consisting of the median pupil dilation value of each participant per group.

5.5.2 Hypothesis 8: Blinking Rate

Hypothesis 8 consists of H_{08} and H_{18}, which question whether there is a significant difference in **blinking rate** in the assembly phase between both groups or not.

A Mann-Whitney U test is also conducted for this data to validate if there are significant differences in blinks per minute between the two groups. The Mann-Whitney U test has shown **no significant effects** between groups ($U = 49.50, p > 0.05, RBC = -0.31, CLES = 0.34$).

This can also be seen in Figure 5.14. The figure shows that, even though the median values differ by 8 blinks, there is no actual significant difference in blinking rate between the two groups in the assembly phase. Therefore, hypothesis H_{08} **cannot be rejected**.

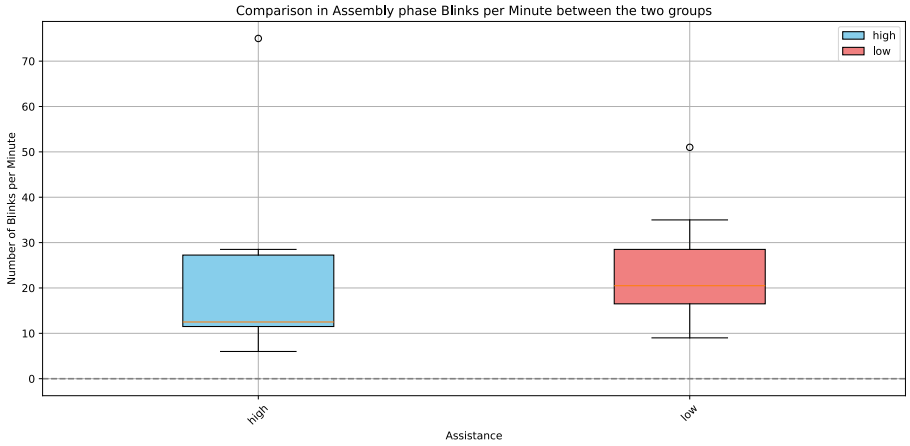


Figure 5.14: The median values of blinks per minute for all participants per assistance group.

5.5.3 Gerenal Participant Behaviour

Besides looking at the main measurements of **pupil dilation and blinking rate**, the duration of a phase can be used to indicate whether a task was easier or harder. For each participant, the duration of both phases is known and thus can be compared. Figure 5.15 shows the duration in seconds for the participants of both groups in the **assembly phase** in a boxplot. The difference in duration for each participant's training and assembly phase can be seen in Table 5.3. Just like for the blinking rate, a Mann-Whitney U test is conducted on the **assembly phase durations**. The Mann-Whitney U test revealed a significant difference in phase duration ($U = 7.0, p < 0.001, RBC = -0.90, CLES = 0.49$). In combination with the data on Figure 5.15, this shows that **low-assistance participants** took a significantly **longer time** to complete the **assembly phase** compared to the **high-assistance** participants. Besides, Table 5.3 shows that, while **high-assistance** participants finished their **assembly phase** quicker compared to their training phase, the opposite happens for 10/12 **low-assistance participants**.

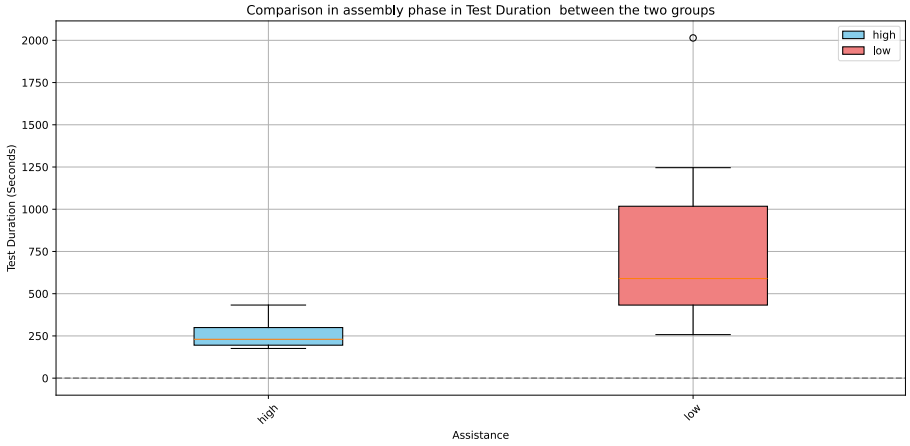


Figure 5.15: A boxplot showing the completion times for both assistance groups.

5.5.4 NASA-TLX Scores

Finally, the subjective data is compared between the two groups to see if the participants perceived the **assembly phase** differently. Figure 5.16 shows the given score to each NASA-TLX

metric per group. The figure shows that for all metrics, **low-assistance** participants generally gave **higher ratings** compared to the **high-assistance** participants. This means that **low-assistance** participants felt like **more mental load** was required in their **assembly phase**, which makes sense since they had to try and remember the correct sticker instead of reading it off the instructions. A Mann-Whitney U test has found that these mental demand ratings differ significantly between the two groups ($U = 10.5, p < 0.001, RBC = -0.85, CLES = 0.07$). They also rated a **higher physical load**, possibly due to the longer durations of their assembly phase. Physical demand ratings are also found to be significantly different between the two groups ($U = 29.5, p < 0.05, RBC = -0.59, CLES = 0.21$). **low-assistance** participants also generally rated **higher temporal demand**, but with much more overlapping data. Besides, from this metric onward, the data is found to be normally distributed. Therefore, a Welch t-test is used. The Welch t-test reveals that the ratings for temporal demand do **not** significantly differ between the two groups($t(21.89) = -1.20, p > 0.05, d = 0.49$). Since there was not really a time limit, this is purely on how the participant set their own goal in the assembly phase before starting. With the **performance metric**, it is shown that **high-assistance** participants mostly rated that they did a reasonable job, with *scores* <= 10. The opposite is true for the **low-assistance** group, who has a mid-range > 7.5 and a lowest rating of 5. This is verified by the Welch t-test revealing significant differences in performance rating ($t(20.10) = -4.48, p < 0.001, d = 1.83$). Also shown is that **low-assistance** participants mostly rated **higher effort** requirements, meaning they had to put in more effort in order to get to their results. This metric is also revealed to show significant differences between the two groups ($t(21.81) = -4.61, p < 0.001, d = 1.88$). Finally, it is shown that **low-assistance** participants generally rated **higher frustration scores**. It is, however, **not** found to significantly differ between the two groups by the Mann-Whitney U test ($t(21.97) = -1.69, p > 0.05, d = 0.69$).

Besides the raw **assembly phase** scores, the subjective difference between the two phases for each group might also be interesting. Figure 5.17 shows the difference in scores for each metric per group. This difference means that for each participant, the **training score is subtracted from the assembly score**. Note that here, all NASA-TLX differences are normally distributed, so a Welch t-test is performed instead of the Mann-Whitney U test. The figure shows that, while **high-assistance** rated lower mental demand in the **assembly phase** compared to the training phase, **low-assistance** members rated higher **mental demand**. So, people who remained assisted stated that the assembly phase is easier, while the opposite is stated by the people who received lower assistance. This is confirmed by the Welch t-test, which revealed the difference is significant ($t(21.51) = -7.13, p < 0.001, d = 2.91$). This means that the scores were similar in the training phase, but the assembly phase's experience after was perceived in a different way between the groups. **physical demand** shows the same trends as mental load, but with differences being closer together. The difference in training and assembly **physical demand** is **not** found to be significantly different between the two groups by the Welch t-test ($t(20.74) = -1.95, p > 0.05, d = 0.79$), indicating the scores were already reasonably different in the training phase. **Temporal demand** is the next metric, which shows somewhat the same trend but with overlapping Quarters. Surprisingly, the Welch t-test reveals that the participants' differences in perception between the two phases do significantly differ ($t(20.54) = -2.33, p < 0.05, d = 0.95$). So although both training and assembly score comparisons were not found to significantly differ, the difference in perception was different. **Performance ratings** generally also show differences between the two groups. High-assistance participants rated their assembly phase to be better, while low-assistance participants mostly rated theirs worse. The Welch t-test reveals this metric to significantly differ between the two groups ($t(21.45) = -3.75, p < 0.01, d = 1.53$), indicating the training scores were similar, but the assembly phase is perceived differently. Also, **Effort ratings** are found to significantly differ between the groups ($t(21.77) = -7.05, p < 0.001, d = 2.88$), which means effort in the training phase was similar but not in the assembly phase. Finally, **frustration scores** are not found to significantly differ between the two groups in the assembly phase, meaning they were similar in the training phase and remained similar in the assembly phase

$(t(17.67) = -0.47, p > 0.05, d = 0.19)$.

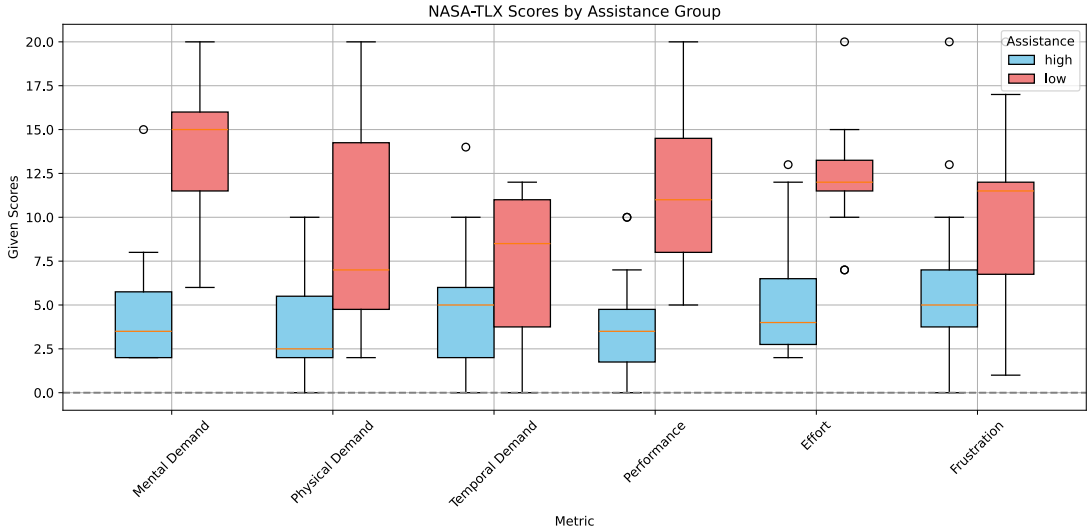


Figure 5.16: The given scores to NASA-TLX metrics for the **assembly phase** between the two groups.

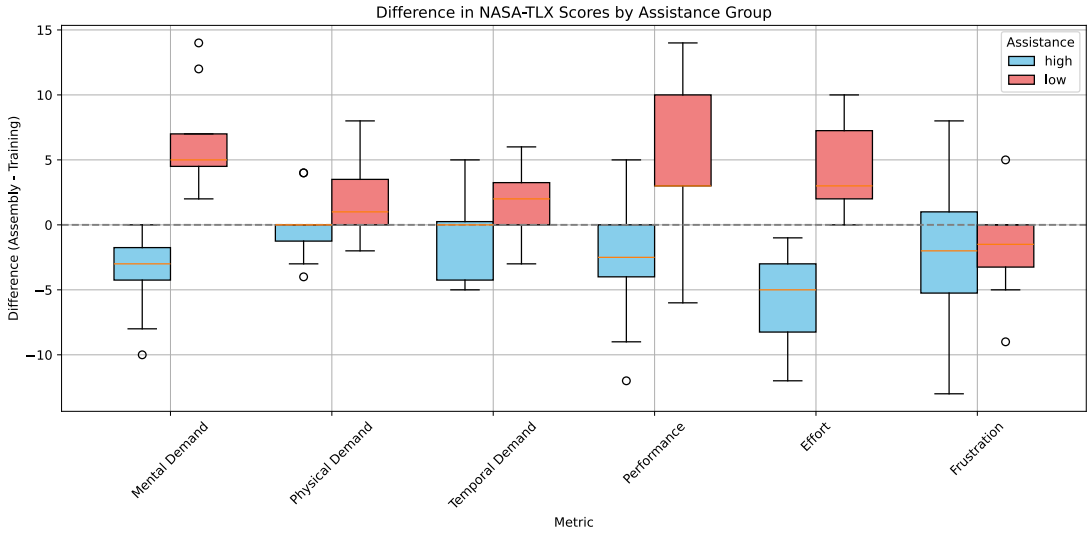


Figure 5.17: The difference in NASA TLX score resulted from subtracting the training phase's score from the assembly phase's score. The boxplot is then the result of the difference for each participant of each group.

Chapter 6

Discussion

For both groups, the changes in pupil dilation and blinking rate were examined. This section will provide some more insight about the findings and interpret them in order to try and answer the research questions.

First of all, this study aims to indicate whether or not cognitive workload can be accurately measured in virtual environments for complex assembly tasks through eye-tracking methods. The study has indicated that the **high-assistance** group revealed **no significant** difference in both **pupil dilation and blinking rate**. This outcome was expected, as their two phases have the exact same level of assistance, thus not requiring participants to put in more mental load. Besides, it is also found that, in contrast to the high-assistance group results, the **low-assistance group** did show a **significant increase** in both **pupil dilation and blinking rate**, which matches the increase of **mental demand scores** in Figure 5.17. This means that the pupil dilation and blinking rate substantially increased when participants had to perform the assembly with a lower form of assistance, while remaining similar when the assistance was unchanged.

Also, when comparing the **assembly phase' pupil dilation** between the two groups, it is found that **low-assistance participants** did show significantly higher pupil dilation compared to the **high-assistance participants**. This indicates that there is a clear sign of pupil dilation increasing when more cognitive workload is required. Besides, it shows that **pupil dilation behavior** matches up with the **mental demand score** shown in Figure 5.17. This shows that a difference between high and low cognitive load can be indicated by pupil dilation measurements.

However, the same can **not** be said for **blinking rate**. Even though the blinking rate rose in the **low-assistance group's assembly phase**, there were no significant differences found when comparing the **assembly phase** between the two groups. This suggests that **blinking rate** was not prone to the cognitive workload, unlike pupil dilation, which did show a significant increase both within and between the groups. Therefore, blinking rate seems to be less of a strong indicator of cognitive load compared to pupil dilation, even though previous studies have identified it to be a relevant indicator [Gur+24]. However, since blinking rate actually generally went up in the **low-assistance group's assembly phase**, there still may be a possibility of blinking rate being connected to mental load, although it cannot be assessed for now.

At this moment, it is still too soon to properly state that cognitive workload can accurately be measured in a complex VR assembly task. The general increase in pupil dilation is present, indicating the possibility of using the metric for indicating cognitive load. However, when looking at the correlations, they do not seem to properly indicate what was expected. For starters, there is no correlation between the participants' self-reported mental demand and their pupil dilation, which is what this study aims to validate. There is, however, one significant correla-

tion, which states that participants who rated more physical demand tended to have smaller pupils. A study by Kuwamizu *et al.* [Kuw+22] suggests that, with more physical demand, pupil dilation should increase, even though it might only be because of arousal. However, Bicalho *et al* [Bic+19] has found more pupil dilation and higher blinking rates in less repetitive tasks compared to more repetitive ones. Since the physical aspect of the task is repetitive, this might be the reason for this correlation. This correlation also does not have any effect on the increase of pupil dilation due to cognitive load, meaning it can exist coherently. Still, even though there does not seem to be much interference, it would be better to do further research on the pupil dilation. Overall, the trend checks out, but on a deeper level, the connection between participants' perceived mental load and their pupil dilation is not in sync. In order to state that pupil dilation can accurately be measured, it would be preferred to have stronger data for these reasons behind the pupil dilation.

In regard to the blinking rate, it may have been influenced by other occurrences rather than mental load, meaning it might have been masked. For example, there is a correlation stating that participants who made more mistakes tended to have a higher blinking rate. This means that lower blinking rates might indeed give a nod towards a higher focus [Ist22]. However, there is also a correlation stating that participants who reported more frustration tended to blink more as well. Thus, frustration might have also played a role in the blinking rate, possibly masking its relation to cognitive load. To add more to it, the blinking rate also correlated with test duration, which tells participants who spent a longer time looking at the VR screen blinked more, although this correlation mostly seems to depend on one outlier and may not be an actual cause.

The study also wanted to answer how factors that interfere with pupil dilation can be eliminated. These interfering factors are based on the first study, which aimed to explore how certain factors influence pupil dilation. Color combinations were found that do not significantly differ in influence. Where possible, colors within the environment were adapted to these non-significant influencing colors. Besides, the skybox and floor were also changed to one solid color that is common on the electrical cabinet itself, decreasing color variation. Finally, a baseline pupil measurement was taken to better normalize the data. The pupil dilation normalization to Z-scores is inspired by the study of the brightness test. Since a significant growth of the pupils was found only when the assistance level dropped, it indicates that the changes made to the virtual environment and the baseline calibration were sufficient in suppressing other influences on pupil dilation. It even worked so well, the black wall calibration process did not even need to be used to correct pupil dilation whenever it would be influenced too much by differences in luminance, giving the changes done to the environment even more credibility.

The final question was whether possible hiccups within an assembly process could be identified through these measurements. When looking at the steps that caused significant differences in pupil dilation, four steps of the process pop up in the list for both groups. These steps have varying results in terms of how they affected pupil dilation in the assembly phase compared to the training phase between the two groups. However, there is one step that increased pupil dilation in the assembly phase for both groups. This could be caused by the sticker of the step having a very similar name to two other stickers. Anyhow, there is a slight indication that this might be possible, but there is no certainty because there is not enough evidence on this subject. To find out if this is truly possible, more research will be needed.

Chapter 7

Conclusions

In this final chapter, the conclusions which are drawn from the gathered data are discussed. Besides, this chapter also reflects upon the subject of this thesis and the learning process that has been gone through when implementing and writing this thesis. Finally, this chapter will provide a critical view of the results.

This study investigates how cognitive load can be measured when performing realistic assembly tasks within a virtual environment, and how these virtual environments should be adapted to reduce interference with external factors. To indicate how the virtual environments should be adapted, a short user study was done, which tested **brightness, color, and depth**. The goal of this study was to investigate the pupil influences on the environment. The study reveals that **pupils dilate differently for different colors**, meaning some colors might cause pupils to shrink and others may cause them to grow. However, there are color combinations that do not significantly differ in effect. When designing a virtual space to measure cognitive load, these colors can be used in the environment to minimize interference with the color effects. The study also revealed that the **virtual depth** should **not** be an issue. When color combinations are used that do **not significantly differ** in pupil dilation effects, the eyes do not respond to the virtual depth. Finally, the test gave a direction towards the possibility of measuring cognitive workload by showing significant differences between the 1-back test and the 2-back test, and the 1-back test and the 3-back test, since these pupil dilation values significantly differed. There were, however, **no significant differences** found between the **two lighting conditions**, which contradicts the paper from which the test was taken. To play it safe, the brightness conditions within the assembly test were kept constant to avoid the pupil dilation being influenced by brightness changes.

After exploring the pupil influences, the main user study started, which had two groups of participants place stickers on an electric cabinet assembly. For both groups, the **training and assembly phases** are compared to see if there were differences in eye behavior. Besides, the **assembly phases** of both groups were compared to reveal whether the difference in assistance had significant effects on participants' eye behavior.

As previously discussed, it is found that pupil dilation did not significantly change when the assistance level remained the same, and that it significantly rose once the level of assistance dropped. This means that it is indeed possible to globally indicate where a higher cognitive load is induced, and thus, pupil dilation can generally be used to indicate a rise in cognitive workload. However, it is still not totally certain whether it is fully linked to cognitive load, since pupil dilation and perceived mental load by participants did not correlate. Although it is **too soon to confidently state that pupil dilation can accurately be measured in complex assembly tasks, there is a strong indication that points towards the possibility**. Future studies should, however, try and find these meaningful correlations to give strong evidence of the true meaning behind these eye behaviors.

Next to pupil dilation, the **blinking rate** was also investigated, but it seemed less prone to cognitive load. Correlations are found, which might indicate that the blinking rate was more related to the perceived frustration of the participant instead of the cognitive load. This means that the blinking rate's relation to cognitive load may have been masked by the frustration participants felt during the tests. As seen by the significant increase within the low-assistance group, it seems like there is still some value left in using blinking rate to indicate cognitive load, but further studies should verify this.

Also, since an increase in **pupil dilation** was found in the **assembly phase of the low-assistance** group, it indicates that the adaptations to the virtual environment sufficiently omitted other influences on the pupil dilation. Thus, as the first user study found, using non-significant color combinations and having a solid baseline calibration did manage to find the expected results overall. This means that **carefully using colors that show no significant effects between each other, making the surrounding one solid common color, and capturing a baseline that includes all colors, including the ones that should significantly affect pupil dilation compared to other colors, is key to indicating cognitive load through eye measurements**. If these steps are considered in the virtual environment, pupil influential metrics can be eliminated enough to indicate higher or lower cognitive load.

For the last question, where it is possible to identify hiccups within the assembly process, there is not enough evidence to say whether this is possible or not. There is one step found that significantly increased pupil dilation, but other than this step, there is nothing else that really supports this. Therefore, for now, **hiccups within an assembly process cannot yet be identified**.

Personally, I think this is a very interesting study that may benefit many businesses aiming to better understand their workers' working behaviors and help them eventually upgrade their working environment by finding mentally demanding steps in the process, which could then lead to a possible solution being compiled. A strong indication of the possibility of measuring cognitive workload for complex assembly tasks in VR is found, but more research is required to really explain is these effects are present for the correct reasons.

During this thesis, I have gained valuable knowledge on multiple aspects. First, I learned how to work with VR applications in Unity, where all the user study material was made. Besides, due to this thesis measuring eye metrics, I also gained some knowledge in eye behaviors, like knowing how they react to color, stress, mental load, and how blinks can indicate focus. Finally, this thesis also made me dive deeper into statistics, on which I had very little knowledge. This is specifically valuable knowledge for future careers.

7.1 Future Work

This study focuses solely upon measuring cognitive workload using eye measuring metrics, such as the pupil's dilation and the number of times a person blinks per minute. However, other metrics could also be used to indicate cognitive workload, such as heart rate, as studied by Abdurrahman *et al.*[Abd+22].

Besides this possible extension, this study also has some limitations. All of the participants in the study are students in computer science at Hasselt University, with a non-varying age and background. Besides, most participants were also males, so females are not greatly represented. The sample size of 12 participants per group could also be expanded to gather even more data. Another limitation could be that participants did the same exact task twice. In order to see if cognitive workload can be measured in complex assembly tasks, it might also be interesting to take a look at other additional assembly tasks, perhaps even in a different working sector. Finally, another limitation that might have had a big effect is that many participants failed to make stickers snap to the electric cabinet. A few participants who had VR experience and

one without any VR experience managed to place down stickers with ease, but many of the participants struggled a lot, which might have created noise in the data.

Although pupil dilation could not be clearly explained by studying its correlation with other metrics, it does show significant effects between the two groups and between the **low-assistance** their phases. As expected, the general pupil dilation did indeed go up, which indicates the foundation for measuring the cognitive workload via pupil dilation is there. Future studies will need to build upon this study and go further to investigate the next step of how this pupil dilation difference can be explained and whether or not it is in fact related to cognitive workload. A big suggestion for this would be to carefully upgrade the **low-assistance** to their assistance. Although seven participants managed to finish with fewer than 20 mistakes, which is still quite a lot, other participants just resorted to trying out every sticker because they could not remember it. Therefore, perhaps not giving any instructions in order to induce as much cognitive load as possible might not have been the best solution.

Bibliography

[Abd+22] Usman Alhaji Abdurrahman et al. "Heart Rate and Pupil Dilation As Reliable Measures of Neuro-Cognitive Load Classification". In: *2022 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC)*. 2022, pp. 1–7. DOI: 10.1109/ASSIC55218.2022.10088296.

[Ahm+23] Mohammad Ahmadi et al. "Comparison of Physiological Cues for Cognitive Load Measures in VR". In: *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. 2023, pp. 837–838. DOI: 10.1109/VRW58643.2023.00261.

[Aku+23] Artem Akulov et al. "Training simulators for crane operators and drivers". In: *Engineering Today* 2 (Jan. 2023), pp. 5–5. DOI: 10.5937/engtoday2300003A.

[Auc18] AucklandEye. *7 Things That Make Your Pupils Change in Size*. https://www.aucklandeye.co.nz/blog/7-things-that-make-your-pupils-change-in-size/. Accessed: 04/01/2025. 2018.

[Bic+19] Lucas Eduardo Antunes Bicalho et al. "Oculomotor behavior and the level of repetition in motor practice: Effects on pupil dilation, eyeblinks and visual scanning". In: *Human Movement Science* 64 (2019), pp. 142–152. ISSN: 0167-9457. DOI: https://doi.org/10.1016/j.humov.2019.02.001. URL: https://www.sciencedirect.com/science/article/pii/S016794571830695X.

[Bio+23] Francesco N. Biondi et al. "Distracted worker: Using pupil size and blink rate to detect cognitive load during manufacturing tasks". In: *Applied Ergonomics* 106 (2023), p. 103867. ISSN: 0003-6870. DOI: https://doi.org/10.1016/j.apergo.2022.103867. URL: https://www.sciencedirect.com/science/article/pii/S0003687022001909.

[CB23] Nairanjana Chowdhury and Chandan Kumar Bhattacharyya. "Measuring The Performance Ability Threshold Of An Individual Under Perceived Stress With Cognitive Load: A Comprehensive Approach". In: *2023 3rd International Conference on Innovative Sustainable Computational Technologies (CISCT)*. 2023, pp. 1–6. DOI: 10.1109/CISCT57197.2023.10351267.

[EHR21] Marie Eckert, Emanuël A. P. Habets, and Olli S. Rummukainen. "Cognitive Load Estimation Based on Pupillometry in Virtual Reality with Uncontrolled Scene Lighting". In: *2021 13th International Conference on Quality of Multimedia Experience (QoMEX)*. 2021, pp. 73–76. DOI: 10.1109/QoMEX51781.2021.9465417.

[El 24] Mohamad El Haj. "Pupillometry as tool to assess cognitive and affective processing in aging". In: *Brain Disorders* 14 (2024), p. 100129. ISSN: 2666-4593. DOI: https://doi.org/10.1016/j.dscb.2024.100129. URL: https://www.sciencedirect.com/science/article/pii/S2666459324000143.

[GSS24] Vanshika Garg, Vaishnavi Singh, and Lav Soni. "Preparing for Space: How Virtual Reality is Revolutionizing Astronaut Training". In: *2024 IEEE International Conference for Women in Innovation, Technology Entrepreneurship (ICWITE)*. 2024, pp. 78–84. DOI: 10.1109/ICWITE59797.2024.10503238.

[Gur+24] Mustafa Can Gursesli et al. "Understanding Game Performance: A Study of Eye Blinking and Pupil Metrics in Matching Pairs Game". In: *2024 IEEE Conference on Games (CoG)*. 2024, pp. 1–6. DOI: 10.1109/CoG60054.2024.10645563.

[GVM22] Ivan Garcia, Emília Villani, and João Mello. "Evaluation of Cognitive Workload in Automated Drilling Processes for Aerospace Structures". In: *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*. 2022, pp. 1–6. DOI: 10.1109/HORA55278.2022.9800063.

[Isk+19] Julie Iskander et al. "Exploring the Effect of Virtual Depth on Pupil Diameter". In: *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. 2019, pp. 1849–1854. DOI: 10.1109/SMC.2019.8913975.

[Ist22] IsthmusEyeCare. *The Purpose of Blinking*. https://isthmuseye.com/patient-care/blog/the-purpose-of-blinking/. Accessed: 27/05/2025. 2022.

[JSL17] Petar Jerčič, Charlotte Sennersten, and Craig Lindley. "The effect of cognitive load on physiological arousal in a decision-making serious game". In: *2017 9th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games)*. 2017, pp. 153–156. DOI: 10.1109/VS-GAMES.2017.8056587.

[Kuw+22] Ryuta Kuwamizu et al. "Pupil-linked arousal with very light exercise: pattern of pupil dilation during graded exercise". In: *The Journal of Physiological Sciences* 72 (Dec. 2022), p. 23. DOI: 10.1186/s12576-022-00849-x.

[Lee+24] Joy Yeonjoo Lee et al. "Measuring Cognitive Load in Virtual Reality Training via Pupillometry". In: *IEEE Transactions on Learning Technologies* 17 (2024), pp. 704–710. DOI: 10.1109/TLT.2023.3326473.

[Li+22] Chunping Li et al. "Cognitive Load Measurement in the Impact of VR Intervention in Learning". In: *2022 International Conference on Advanced Learning Technologies (ICALT)*. 2022, pp. 325–329. DOI: 10.1109/ICALT55010.2022.00103.

[NM24] Regina Nádas and György Molnár. "The Relationship Between Attention, Memory and Pupillometry". In: *2024 IEEE 7th International Conference and Workshop Óbuda on Electrical and Power Engineering (CANDO-EPE)*. 2024, pp. 317–322. DOI: 10.1109/CANDO-EPE65072.2024.10772941.

[Ore+12] Mike Oren et al. "Puzzle assembly training: Real world vs. virtual environment". In: *2012 IEEE Virtual Reality Workshops (VRW)*. 2012, pp. 27–30. DOI: 10.1109/VR.2012.6180873.

[PKV21] Giovanni Pignoni, Sashidharan Komandur, and Frode Volden. "Accounting for Effects of Variation in Luminance in Pupillometry for Field Measurements of Cognitive Workload". In: *IEEE Sensors Journal* 21.5 (2021), pp. 6393–6400. DOI: 10.1109/JSEN.2020.3038291.

[PS03] Marc Pomplun and Sindhura Sunkara. "Pupil dilation as an indicator of cognitive workload in human-computer interaction". In: *Proceedings of the International Conference on HCI* (Jan. 2003).

[Sev+22] Natalia Sevcenko et al. "Theory-based approach for assessing cognitive load during time-critical resource-managing human–computer interactions: an eye-tracking study". In: *Journal on Multimodal User Interfaces* 17 (Nov. 2022), pp. 1–19. DOI: 10.1007/s12193-022-00398-y.

[Sha+12] Shakil Shaikh et al. "Investigating the effects of physical and cognitive demands on the quality of performance under different pacing levels". In: *Work* 41 (Jan. 2012), pp. 1625–1631. DOI: 10.3233/WOR-2012-0363-1625.

[SS24] Simon Schwerd and Axel Schulte. "Evaluating Blink Rate as a Dynamic Indicator of Mental Workload in a Flight Simulator". In: *VISIGRAPP*. 2024. URL: https://api.semanticscholar.org/CorpusID:268234222.

[Sul+20] Shahida binti Sulaiman et al. "Virtual Reality Training and Skill Enhancement for Offshore Workers". In: *2020 International Conference on Computational Intelligence (ICCI)*. 2020, pp. 287–292. DOI: 10.1109/ICCI51257.2020.9247819.

[Sur+20] Maman Suryaman et al. "Tailoring The Certified of Electric Engineering for Power Plant: Development Training Model by Using 3D VR Simulator Modification for Professional Engineer Case Study of Universitas Singaperbangsa Karawang". In: *2020 3rd International Conference on Mechanical, Electronics, Computer, and Industrial Technology (MECnIT)*. 2020, pp. 267–270. DOI: 10.1109/MECnIT48290.2020.9166674.

[Yu+23] Haiyang Yu et al. "Effects of cognitive load on human cognitive reliability". In: *2023 11th International Conference on Information Systems and Computing Technology (ISCTech)*. 2023, pp. 220–224. DOI: 10.1109/ISCTech60480.2023.00047.

Appendix

High-assistance

Participant	**Seconds looked at resource interface per minute of test**													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Training														
P3	19	19	24	12	17	/	/	/	/	/	/	/	/	/
P4	26	8	13	16	29	30	25	16	/	/	/	/	/	/
P5	20	29	25	39	19	28	15	0	0	14	7	3	24	29
P6	21	22	14	25	15	4	15	12	9	18	4	/	/	/
P7	31	10	11	20	13	11	17	9	/	/	/	/	/	/
P8	42	23	32	4	13	10	27	26	21	30	/	/	/	/
P9	16	17	23	33	7	28	23	26	24	/	/	/	/	/
P10	11	28	19	14	24	18	12	16	27	/	/	/	/	/
P11	31	27	20	25	26	11	/	/	/	/	/	/	/	/
P12	6	14	26	14	25	14	20	/	/	/	/	/	/	/
P13	22	30	25	27	/	/	/	/	/	/	/	/	/	/
Assembly														
P3	17	17	27	10	/	/	/	/	/	/	/	/	/	/
P4	34	15	23	8	/	/	/	/	/	/	/	/	/	/
P5	38	33	17	36	8	17	11	/	/	/	/	/	/	/
P6	24	23	16	18	13	20	/	/	/	/	/	/	/	/
P7	20	15	19	15	5	9	15	4	/	/	/	/	/	/
P8	41	24	31	33	21	/	/	/	/	/	/	/	/	/
P9	19	36	23	11	4	/	/	/	/	/	/	/	/	/
P10	36	11	23	19	13	/	/	/	/	/	/	/	/	/
P11	33	33	28	15	/	/	/	/	/	/	/	/	/	/
P12	29	25	33	6	/	/	/	/	/	/	/	/	/	/
P13	28	21	21	5	/	/	/	/	/	/	/	/	/	/

Table 1: The number of seconds spent looking at the resource interface per minute of the test (rounded to the nearest second) in the high-assistance group. This is both for the training and the assembly test. The 15th minute is removed since only one participant had a 15th minute, but the participant ended 3 seconds after the 15th minute started.

Participant	**Amount of blinks per minute of test**													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Training														
P2	11	13	1	5	9	0	/	/	/	/	/	/	/	/
P3	23	25	16	17	15	/	/	/	/	/	/	/	/	/
P4	51	38	41	39	29	38	49	23	/	/	/	/	/	/
P5	1	18	6	40	64	142	21	22	9	13	18	20	8	33
P6	22	17	12	19	23	24	27	51	40	32	9	/	/	/
P7	38	33	27	28	36	27	34	9	/	/	/	/	/	/
P8	17	13	9	13	17	9	16	12	6	9	/	/	/	/
P9	19	17	23	28	17	39	17	24	26	/	/	/	/	/
P10	17	7	18	11	12	12	30	14	4	/	/	/	/	/
P11	10	10	12	8	13	3	/	/	/	/	/	/	/	/
P12	6	12	20	7	13	12	6	/	/	/	/	/	/	/
P13	10	12	19	12	/	/	/	/	/	/	/	/	/	/
Assembly														
P2	4	7	6	/	/	/	/	/	/	/	/	/	/	/
P3	11	15	20	9	/	/	/	/	/	/	/	/	/	/
P4	32	22	40	8	/	/	/	/	/	/	/	/	/	/
P5	38	35	96	29	75	137	161	/	/	/	/	/	/	/
P6	21	12	3	7	12	12	/	/	/	/	/	/	/	/
P7	50	28	45	27	29	25	47	11	/	/	/	/	/	/
P8	10	17	9	15	6	/	/	/	/	/	/	/	/	/
P9	14	33	34	28	2	/	/	/	/	/	/	/	/	/
P10	16	9	19	12	5	/	/	/	/	/	/	/	/	/
P11	6	6	8	6	/	/	/	/	/	/	/	/	/	/
P12	14	10	23	2	/	/	/	/	/	/	/	/	/	/
P13	17	10	16	2	/	/	/	/	/	/	/	/	/	/

Table 2: The real number of blinks per minute of the test. This is both for the training and the assembly test in the high-assistance group. If a participant's time was less than a column's minute value, the column indicates a "/". The 15th minute is removed since only one participant had a 15th minute, but the participant ended 3 seconds after the 15th minute started.

Participant	Median pupil dilation change in Z-score per minute													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Training														
P2	-2.182	-1.933	-0.681	-0.348	-1.233	0.911	/	/	/	/	/	/	/	/
P3	-1.945	-3.395	-3.072	-3.259	-2.626	/	/	/	/	/	/	/	/	/
P4	-1.971	-2.104	-1.648	-1.895	-1.427	-1.311	-1.226	-0.941	/	/	/	/	/	/
P5	-4.295	-5.794	-5.001	-6.746	-6.353	-5.854	-4.418	-6.611	-7.703	-6.02	-5.671	-5.834	-6.935	-5.466
P6	-0.567	-1.873	-2.077	-1.842	-2.994	-1.306	-0.824	-1.733	-2.127	1.485	1.476	/	/	/
P7	-3.207	-2.938	-3.478	-3.077	-4.551	-3.591	-3.745	-2.405	/	/	/	/	/	/
P8	0.021	-1.13	-1.243	-1.114	-1.243	-0.867	-0.529	-0.494	-0.876	-0.667	/	/	/	/
P9	-0.661	-0.522	-0.579	-0.686	-1.118	-0.926	-1.235	-1.251	-0.553	/	/	/	/	/
P10	0.066	-0.892	-0.973	-0.469	-0.417	-0.759	-0.328	-0.693	-0.595	/	/	/	/	/
P11	-2.201	-3.561	-3.141	-2.723	-2.585	-3.385	/	/	/	/	/	/	/	/
P12	-5.821	-5.681	-5.258	-6.226	-5.122	-4.084	-4.683	/	/	/	/	/	/	/
P13	-3.21	-4.78	-4.111	-4.048	/	/	/	/	/	/	/	/	/	/
Assembly														
P2	-1.343	-1.640	-0.970	/	/	/	/	/	/	/	/	/	/	/
P3	-3.181	-4.251	-4.429	-3.775	/	/	/	/	/	/	/	/	/	/
P4	-1.37	-1.31	-0.907	-0.482	/	/	/	/	/	/	/	/	/	/
P5	-4.012	-5.901	-6.386	-5.280	-6.166	-6.114	-5.588	/	/	/	/	/	/	/
P6	-1.363	-3.656	-2.095	-2.569	-3.352	-1.222	/	/	/	/	/	/	/	/
P7	-7.237	-7.396	-7.418	-7.614	-6.578	-7.522	-6.944	-5.967	/	/	/	/	/	/
P8	-0.733	-1.600	-1.184	-1.271	-0.529	/	/	/	/	/	/	/	/	/
P9	-1.270	-0.905	-1.178	-1.003	-0.802	/	/	/	/	/	/	/	/	/
P10	-1.141	-1.613	-1.453	-1.035	-1.097	/	/	/	/	/	/	/	/	/
P11	-2.126	-2.828	-2.915	-2.075	/	/	/	/	/	/	/	/	/	/
P12	-4.571	-5.534	-5.112	-3.991	/	/	/	/	/	/	/	/	/	/
P13	-4.090	-4.435	-5.117	-4.059	/	/	/	/	/	/	/	/	/	/

Table 3: The median pupil dilation change in Z-score per minute of the test (rounded to 3 digits behind the comma) for the high-assistance group. This is both for the training and the assembly test. The 15th minute is removed since only one participant had a 15th minute, but the participant ended 3 seconds after the 15th minute started. The dilation change is based on the Z-score of change for each recorded pupil size compared to the median pupil size during calibration.

Low-assistance

Participant	**Seconds looked at resource interface per minute of test**																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Training																	
P14	20	30	33	30	27	21	/	/	/	/	/	/	/	/	/	/	/
P15	20	21	26	10	17	30	20	/	/	/	/	/	/	/	/	/	/
P16	38	35	22	30	31	32	16	/	/	/	/	/	/	/	/	/	/
P17	18	24	29	31	32	16	31	34	/	/	/	/	/	/	/	/	/
P18	35	31	48	15	41	39	0	/	/	/	/	/	/	/	/	/	/
P19	25	24	19	3	4	28	8	/	/	/	/	/	/	/	/	/	/
P20	17	17	31	25	29	11	27	0	9	35	11	24	24	/	/	/	/
P21	16	28	17	24	27	12	33	5	/	/	/	/	/	/	/	/	/
P22	8	3	19	14	24	19	8	7	12	6	16	/	/	/	/	/	/
P23	24	28	17	32	18	/	/	/	/	/	/	/	/	/	/	/	/
P24	15	6	14	4	7	8	14	5	21	25	3	/	/	/	/	/	/
P25	27	29	24	17	18	25	21	0	/	/	/	/	/	/	/	/	/
Assembly	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
P14	20	31	29	19	13	39	35	34	/	/	/	/	/	/	/	/	/
P15	38	40	27	14	44	27	27	/	/	/	/	/	/	/	/	/	/
P16	37	28	21	31	16	13	44	14	24	27	30	19	36	19	/	/	/
P17	31	27	34	30	22	22	14	24	22	18	30	14	14	16	21	23	25
P18	31	33	34	30	30	28	23	11	/	/	/	/	/	/	/	/	/
P19	38	31	38	25	11	/	/	/	/	/	/	/	/	/	/	/	/
P20	33	30	39	28	25	19	13	30	30	19	15	25	24	34	27	34	30
P21	38	33	22	28	25	14	26	3	/	/	/	/	/	/	/	/	/
P22	20	20	22	34	16	11	12	48	14	25	29	39	43	29	38	24	16
P23	28	13	27	14	14	12	22	20	/	/	/	/	/	/	/	/	/
P24	30	15	23	18	20	18	11	37	22	11	15	16	/	/	/	/	/
P25	16	26	53	26	35	17	21	38	19	22	19	21	25	11	22	17	8
Assembly	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
P17	17	25	30	40	/	/	/	/	/	/	/	/	/	/	/	/	/
P20	7	17	28	28	16	27	31	22	13	17	16	22	21	11	23	9	17
P22	10	30	16	/	/	/	/	/	/	/	/	/	/	/	/	/	/

Table 4: The number of seconds spent looking at the resource interface per minute of the test (rounded to the nearest second) in the low-assistance group. This is both for the training and the assembly test. The assembly phase is extended at the bottom for the three participants who exceed 17 minutes.

Participant	**Amount of blinks per minute of test**																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Training																	
P14	28	16	24	23	16	13	/	/	/	/	/	/	/	/	/	/	/
P15	17	20	19	15	31	20	12	/	/	/	/	/	/	/	/	/	/
P16	9	12	9	7	19	12	7	/	/	/	/	/	/	/	/	/	/
P17	18	14	18	17	24	28	20	9	/	/	/	/	/	/	/	/	/
P18	10	23	24	21	27	21	0	/	/	/	/	/	/	/	/	/	/
P19	11	7	5	9	3	10	2	/	/	/	/	/	/	/	/	/	/
P20	42	37	32	58	52	48	55	45	58	20	61	58	24	/	/	/	/
P21	12	15	11	17	12	17	9	2	/	/	/	/	/	/	/	/	/
P22	11	4	10	11	21	15	5	15	26	19	20	22	/	/	/	/	/
P23	20	28	36	24	18	/	/	/	/	/	/	/	/	/	/	/	/
P24	11	12	19	4	7	10	18	7	21	19	0	/	/	/	/	/	/
P25	19	5	11	19	13	7	6	0	/	/	/	/	/	/	/	/	/
Assembly	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
P14	22	26	30	29	28	19	18	29	/	/	/	/	/	/	/	/	/
P15	18	16	14	12	11	16	15	/	/	/	/	/	/	/	/	/	/
P16	101	8	11	29	24	38	26	22	39	40	54	37	68	23	/	/	/
P17	27	24	13	22	20	21	19	26	30	20	21	45	36	23	24	15	9
P18	22	22	22	19	18	17	21	8	/	/	/	/	/	/	/	/	/
P19	10	11	7	9	4	/	/	/	/	/	/	/	/	/	/	/	/
P20	53	58	41	26	41	51	82	56	84	52	58	34	33	39	31	46	53
P21	15	30	17	17	19	17	12	0	/	/	/	/	/	/	/	/	/
P22	11	8	10	18	12	11	8	14	16	21	19	14	16	8	3	12	6
P23	36	29	40	44	34	29	22	37	/	/	/	/	/	/	/	/	/
P24	35	16	18	30	9	36	20	26	26	11	25	19	/	/	/	/	/
P25	27	24	17	14	23	30	12	16	10	10	18	12	42	23	18	23	1
Assembly	18	19	20	21	22	23	4	25	26	27	28	29	30	31	32	33	34
P17	30	14	18	18	/	/	/	/	/	/	/	/	/	/	/	/	/
P20	71	46	109	41	119	60	40	51	37	74	52	41	35	65	40	109	29
P22	19	10	5	/	/	/	/	/	/	/	/	/	/	/	/	/	/

Table 5: The real number of blinks per minute of the test. This is both for the training and the assembly test in the low-assistance group. If a participant's time was less than a column's minute value, the column indicates a "/". The assembly phase is extended at the bottom for the three participants who exceeded 17 minutes.

Participant	**Median pupil dilation change in Z-score per minute**																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Training																	
P14	-11.309	-13.548	-11.372	-11.694	-12.832	-11.715	/	/	/	/	/	/	/	/	/	/	/
P15	1.984	-0.517	0.008	-1.279	1.148	1.561	1.71	/	/	/	/	/	/	/	/	/	/
P16	-0.748	-1.44	-1.216	-0.924	-0.374	-0.97	-0.721	/	/	/	/	/	/	/	/	/	/
P17	-1.328	-1.716	-2.379	-2.428	-2.153	-2.394	-2.281	-1.969	/	/	/	/	/	/	/	/	/
P18	-2.025	-2.97	-2.479	-2.918	-2.058	-2.26	-3.176	/	/	/	/	/	/	/	/	/	/
P19	-0.631	-1.116	-0.627	-1.327	-0.839	-1.306	-0.675	/	/	/	/	/	/	/	/	/	/
P20	-0.541	-0.946	-1.201	-0.961	-0.793	-1.101	-0.824	-0.751	-0.856	-0.579	-0.482	-0.576	-0.425	/	/	/	/
P21	-1.943	-1.767	-3.521	-1.485	-3.121	-3.205	-2.871	-2.356	/	/	/	/	/	/	/	/	/
P22	-0.763	-1.495	-1.188	-1.944	-2.47	-1.518	-1.475	-0.152	-1.074	-1.781	-1.81	-0.116	/	/	/	/	/
P23	-5.29	-7.135	-6.455	-5.513	-5.352	/	/	/	/	/	/	/	/	/	/	/	/
P24	0.807	0.866	1.251	0.912	0.633	0.875	1.061	1.108	0.94	0.939	1.294	/	/	/	/	/	/
P25	-1.522	-2.978	-1.897	-1.684	-1.372	-1.25	-0.687	-1.609	/	/	/	/	/	/	/	/	/
Assembly	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
P14	-9.92	-10.286	-11.237	-13.169	-13.78	-11.956	-10.102	-11.161	/	/	/	/	/	/	/	/	/
P15	4.138	2.714	2.226	1.597	3.801	2.911	3.859	/	/	/	/	/	/	/	/	/	/
P16	-0.362	-0.8	-0.446	-0.586	-0.503	-0.783	-0.647	-0.486	-0.63	-0.618	-0.136	-0.462	-0.56	-0.518	/	/	/
P17	0.174	-1.772	-1.704	-1.273	-1.564	-1.1	-1.324	-1.335	-1.079	-1.318	-1.168	-0.829	-1.296	-0.978	-1.348	-1.403	-1.081
P18	-2.029	-2.24	-2.505	-2.698	-2.537	-2.83	-2.081	-3.212	/	/	/	/	/	/	/	/	/
P19	-0.106	-0.815	-0.163	-0.322	-0.443	/	/	/	/	/	/	/	/	/	/	/	/
P20	-0.246	-0.858	-1.039	-1.284	-1.12	-0.95	-0.906	-0.645	-0.845	-0.862	-0.602	-0.889	-0.951	-0.755	-0.744	-0.554	-0.788
P21	-1.309	-2.458	-2.308	-2.792	-1.889	-0.04	-2.466	-3.475	/	/	/	/	/	/	/	/	/
P22	-1.766	-2.751	-2.076	-1.63	-1.365	-1.25	-1.611	0.295	-1.616	-1.006	-0.205	0.674	-0.672	-1.39	-0.736	-0.316	0.331
P23	-2.738	-3.993	-3.628	-3.649	-2.591	-3.031	-2.384	-2.081	/	/	/	/	/	/	/	/	/
P24	1.011	1.182	1.439	1.779	1.638	1.395	1.389	1.41	0.924	0.976	1.715	1.282	/	/	/	/	/
P25	-2.076	-2.047	-2.717	-2.666	-2.692	-2.753	-2.609	-2.44	-2.538	-2.796	-3.261	-3.358	-3.264	-2.115	-1.955	-1.554	-1.486
Assembly	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
P17	-0.543	0.053	-0.707	-1.572	/	/	/	/	/	/	/	/	/	/	/	/	/
P20	-0.922	-0.472	-0.786	-0.743	-0.413	-0.361	-0.773	-0.802	-0.225	-0.511	-0.807	-0.664	-0.653	-0.754	-0.761	-0.928	-0.001
P22	-0.388	-0.591	0.153	/	/	/	/	/	/	/	/	/	/	/	/	/	/

Table 6: The median pupil dilation change in Z-score per minute of the test (rounded to 3 digits behind the comma) for the low-assistance group. This is both for the training and the assembly test. The dilation change is based on the Z-score of change for each recorded pupil size compared to the median pupil size during calibration. The assembly phase is extended at the bottom for the three participants who exceed 17 minutes.