Usability Across Connected Devices

A laboratory experiment using eye tracking to research the visual distribution and usability across connected devices.

Filip Gustafsson

Gustav Larsson

Computer and System Sciences
Bachelor Thesis 15 Credits
Visual Distribution Across Devices
Interaction design 180 Credits
Spring term 2016
Supervisor: Preben Hansen

Stockholms universitet

Usability Across Connected Devices

A laboratory experiment using eye tracking to research the visual distribution and usability across connected devices

Filip Gustafsson
Gustav Larsson

Abstract

With a steady growth of IoT devices, there's currently a call for papers in the field of Human-Computer Interaction to learn more about how collocated interaction works across connected devices that use the novel technologies that IoT has introduced. One great example of a novel IoT solution is the Philips Hue. Philips Hue are light bulbs that can be controlled using a smartphone, no matter the location of the user. But how do users distribute their visual attention between these connected devices, the smartphone and the light bulbs, when they want to control their lights? Is there a difference in visual attention given by a user depending on if they are by the light bulbs or not, and if so, does it affect the usability of the Philips Hue system as a whole? A laboratory experiment was conducted to explore these themes and answer the research question; *In what ways do the usability of an edge device differ if the user gets visual feedback from the actuator when interacting with the edge device?* A between group design approach with a sample size of ten people per group was applied to compare the difference between having visual access to the light bulbs (experimental group) versus not having visual access (control group). Variables to measure the visual attention given by users was acquired through the use of eye tracking glasses that the participants wore throughout the experiment. Usability was divided into the three variables of effectiveness, efficiency, and satisfaction. Effectiveness and efficiency was measured by letting the participants complete a set of tasks, where each task was measured in a completion rate of pass or fail and completion time was measured in seconds, respectively. The satisfaction was measured using the System Usability Scale, a well-established questionnaire to evaluate the usability of systems. The variables were later analyzed using SPSS to compare the means, standard deviations, probability and effect sizes relevant to help answer the research question. The overall results showed that if the light bulbs were present, the user distributed their visual attention equally between the lightbulbs and the smartphone app. The results also showed that the control group had a greater score of satisfaction and efficiency, while there wasn't a significant difference in effectiveness between the groups. Both effectiveness and efficiency, however, had a rather low probability score which implies that the sample size should have been larger in order to better generalize and validate the results.

Keywords

Internet of Things, Connected Devices, Collocated Interaction, Usability, Visual Distribution, Experiment, Eye tracking, System Usability Scale, Philips Hue, Tobii Pro Glasses 2

List of Content

List of Figures

List of Tables

Acronyms

IoT - Internet of Things
M2M - Machine to Machine
RFID - Radio Frequency Identification
UX - User Experience
UI - User Interface
IEEE - The Institute of Electrical and Electronics Engineers
ISO - The International Organization for Standardization
ETSI - The European Telecommunications Standards Institute
SUS - System Usability Scale
HCI - Human Computer Interaction
AOI - Area of Interest
SPSS - Statistical Package for the Social Sciences
M - Means
SD - Standard Deviation
N - Sample Size
df - Degrees of freedom
CI - Confidence Interval
p – Probability
CTA – Concurrent Think Aloud
CSCW – Computer Supported Cooperative Work
QUIS – Questionnaire for User Interaction Satisfacton
SUMI – Software Usability Measurement Inventory
PSSUQ - Post-Study System Usability Questionnaire
CSUQ – Computer System Usability Questionnaire
DSLR – Digital Single-Lens Reflex Camera

Dictionary

IoT system - A system of multiple IoT devices, i.e. sensors, actuators, gateways and edge devices.

Actuator - The Device being controlled.
In this study: the Philips Hue light bulbs serve as the actuators in the IoT system.

Edge Device - The device on the other end of the actuator, that a user interacts with to control the actuator.
In this study: the smartphone app serves as the user controlled edge device in the IoT system.

1. Introduction

1.1 Background

"*Internet is the future*", a claim that has been true for the past two decades and is still going strong[1][2][3]. Within five years, Ericsson AB estimates 28 billion devices will be connected to the Internet[4], Cisco has it bets on 50 billion[5], while Juniper Research Ltd calculates a growth to 38 billion devices[6]. No matter what the future holds, these connected devices have had a traction big enough to grant them a label of their own, known as the Internet of Things[7], abbreviated as IoT.

Research made within IoT have been struggling to find common ground for the term IoT itself. Professor Theo Kanter who currently specializes in IoT emphasize a need for standardizing the area, and once defined IoT as following[8]:

> *"[IoT] It's about smart sensors with wireless connectivity to the network. Internet as we know it today will still be here, but what is new is that it be expanded with more connected artefacts as a way of integrating the real and digital world."* – Theo Kanter

The term IoT was arguably first coined by Kevin Ashton back in 1999. It was coined to describe a change in which radio frequency identification (RFID) combined with sensors could be used to transfer data from machine to machine (M2M) without the need of human interaction. This was a change Ashton himself saw had the potential to create a paradigm shift where computers could learn about the world using sensors of their own rather than having humans inputting data.[9]

A lot has happened since. During these past 17 years, following the term's commercial success and expansion to include a wide range of devices, many believe that the term IoT has failed to evolve enough and convey the entirety of its implications. An IoT system is often described in terms of an edge device that sends data to an actuator through a gateway in either an internal or external network. The exclusion of users in the description of IoT is a critique that emphasizes the importance of learning more about the relationship between users and connected devices.[7]

---

[1] Gates, B. 1999. Business @ the speed of thought. *1st. Pearson Education Limited.*
[2] Wolf, G. 1996. Steve Jobs: The Next Insanely Great Thing. www.*wired.com.*
[3] Page, L. 2014. Where's Google going next? *www.ted.com.*
[4] Ericsson AB. 2015. Ericsson Mobility Report: Internet of Things to overtake mobile phones by 2018. *www.ericsson.com.*
[5] Evans, D. 2011. The Internet of Things: How the Next Evolution of the Internet Is Changing Everything. www.*cisco.com*
[6] Juniper Research Ltd. 2015. 'Internet of Things' Connected Devices to Almost Triple to Over 38 Billion Units by 2020. *www.juniperresearch.com.*
[7] Rowland, C. Goodman, E. Charlier, M. Light, A. Lui, A. 2015. Designing Connected Products: UX for the Consumer Internet of Things. *O'Reilly Media, Inc.*
[8] Kanter, T. n.d. Smart Service For Future Internet. *DSV Research, Stockholm University.*
[9] Ashton, K. 2009. That 'Internet of Things' Thing. *www.rfidjournal.com.*

This critique is important to recognize for multiple reasons. As more and more devices join the ubiquitous rank of being connected in order to enhance their interusability, the IoT label may in time lose its purpose completely. Connected devices will allegedly be considered, not only part of the mainstream Internet but, a natural part of our everyday society[7]. Furthermore, not considering users' roles when designing connected devices precludes any reason for a designer to consider creating a cohesive user experience (UX), if there's as allegedly claimed; no users involved in IoT.

Yet there is much to learn. As the technology matures, the latter critique could be explained as to why designers are still at an infancy level of understanding what makes a compelling UX across multiple connected devices[7].

1.2 Problem

Figure 1. Commercial for the X10 Powerhouse

Connected devices has been around for a while but it's still evolving[10]. The first widely advertised system that used connected devices was the X10 Powerhouse that was launched all the way back in 1986[11] (See figure 1).

While the technology might seem ancient compared to today's standards, many of the challenges that the X10 Powerhouse faced still remain for many of the IoT devices that are being marketed to consumers today. Convincing the mass-market the value of breaking their current everyday habits in favor of solutions using IoT devices is hard, because predicting human behaviors and needs are hard if the users can't recognize the benefit of breaking their current habits[12]. For example, having a system that uses geographical proximity to turn on the lights in a home when the resident gets close may sound like a clear boon in theory, while in practice it may turn out that the resident doesn't always need the lights to turn on when they get home.

Designing IoT devices that fail to take account for the needs of humans that are part of the IoT network and recognize how the user interacts with the system will most likely result in an incohesive UX[11]. Emphasis being on most likely, because the fact is that there's still largely a gap of knowledge regarding the user's role within the definition of IoT and the importance of usability for connected devices[12].

Designers and researchers still know little regarding how users perceive the usability of connected products. To define and determine if certain devices are to be considered IoT or not is less important, assuming the term will be rendered redundant in the near future regardless.[12] By learning more about the usability of connected products, i.e. how collocated interactions has an impact on usability or how

[10] Minerva, R. Biru, A. Rotondi, D. 2015. Towards a definition of the Internet of Things (IoT). *www.ieee.org.*
[11] Commodore. 1986. The X-10 Powerhouse does everything but put out the cat. *www.commodore.ca*
[12] Rowland, C. Goodman, E. Charlier, M. Light, A. Lui, A. 2015. Designing Connected Products: UX for the Consumer Internet of Things. *O'Reilly Media, Inc.*

visual distribution across connected devices affects usability, future designers could opt for greater usability and improved UX for connected products.

As late as of May this year, the HCI Journal made a call for papers for a special issue on the subject of collocated interaction. They propose the contributions needed within the field of Human-Computer Interaction (HCI) should relate to three points:[13]

- Studies of novel collocated interaction settings (including ‘everyday’ settings and multi-device ecologies);
- Design and study of interaction techniques and systems supporting collocated interaction (including cross-device interaction, spatial interaction, mixed-presence, and groupware systems);
- Discussion of methodologies and theoretical approaches to study and design collocated interaction.

This clearly implies there’s a current need to research the concept of collocated interactions across devices.

1.3 Research Question

In what ways do the usability of an edge device differ if the user gets visual feedback from the actuator when interacting with the edge device?

Usability can be measured in the quantitative terms of effectiveness, efficiency and satisfaction. The goal of this study was to compare the usability measurements from two different scenarios. In one scenario the light bulbs (actuators) and smartphone app (edge device) were both present, while in the other scenario the light bulbs (actuators) were not present - only the smartphone app (edge device). The second scenario simulated a situation where the user had no visual access to the light bulbs (actuators). The IoT system in use to evaluate the usability was the Philips Hue System (see figure 2).

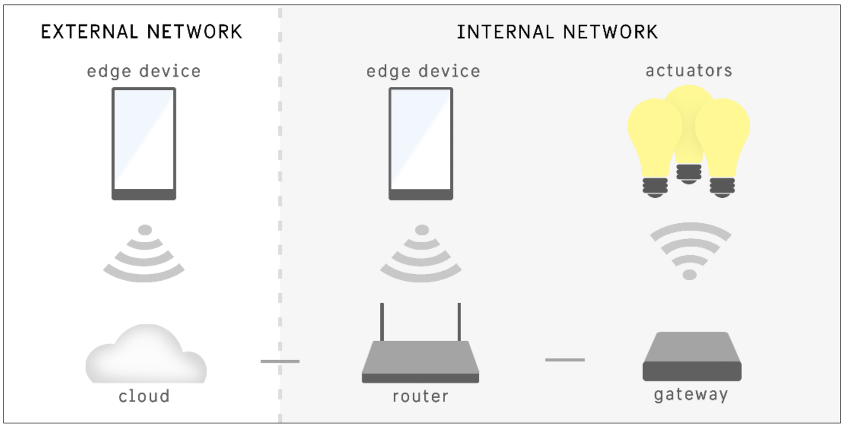

Figure 2. The Philips Hue consist of a gateway connected to a router, an Internet service, multiple actuators (light bulbs) and an edge device (smartphone app), making it representative of a complete IoT system

[13] Fisher, J. 2016. HCI Journal Special Issue on Collocated Interaction: New Challenges in ‘same time, same place’ Research. *CSCW 16´ San Francisco, USA.*

The Philips Hue uses the familiar concept of turning on and off lights, only altering the way in which the lights are being interacted with through the capabilities of IoT. This makes it easier for participants previously unfamiliar with IoT to better understand the premise of the chosen IoT system. The Philips Hue System is marketed as a product that can be used either home or away from home[14], making it ideal to evaluate the concept of collocated interaction across devices and help answer the research question. As it stands today, no distinguished studies have been published regarding how the usability of an edge device changes depending on if the user has visual access to the actuators or not.

1.4 Hypothesis

People who use a smartphone app to interact with a light bulb (actuator) devotes less visual attention to the smartphone app (edge device) in favor of the light bulb (actuator) if present, which enhances the usability of the IoT system in terms of effectiveness, efficiency and satisfaction.

[14] Philips. 2012. Introducing Philips hue: the world's smartest LED bulb, marking a new era in home lighting. *www.philips.com.*

2. Extended Background

2.1 Internet of things

As recent as last year, Institute of Electrical and Electronics Engineers (IEEE) published a comprehensive document that gave an all-inclusive summary of how IoT has been defined so far in order to establish a definition of both the term IoT and its basic architectural requirements for implementation. IEEE describes IoT as a network of items, each embedded with sensors, which are connected to the Internet.[15] In this paper, however, there's no intention to suggest a definition to rule them all, but rather summarize current perspectives and provide a definition that's sufficient enough to explore and discuss ways in which users interact with connected devices.

2.1.1 The History of Internet of Things

The concept of IoT dates back to the second world war. It all started with the use of RFID on returning airplanes that couldn't be identified by the ground control through the radar-system. This was solved with airplanes performing a spin midair which made the radar waves reflect back in an alternative way, giving them a unique identifier. RFID today works in the same basic concept. A signal is sent to a transponder which wakes up and either reflects back a signal if it's a passive system, or broadcast a signal if it's an active system.[16]

In 1973 Charles Walton received a patent for a system which used RFID to unlock a door without a traditional key[16]. That technology is the first dated technology that closely relates to the way RFID is used in IoT today.

The term IoT was coined by Kevin Ashton, by the time the executive director of Auto-ID Center, in the year 1999. He described IoT as a concept of using RFID-tags in a supply chain to track individual objects.[17] This definition of IoT changed when Sanjay Sarma and David Brock linked the RFID-tags into a network that connected any object with a RFID-tag to the Internet. Manufacturers could e.g. automatically let a business partner know when a shipment left the factory, which vastly improved the communication between both parties by making the feedback direct and automatized.[16]

[15] Minerva, R. Biru A. & Rotondi, D. 2015. Towards a definition of the Internet of Things (IoT). *www.ieee.com*
[16] Roberti, M. 2005. The History of RFID Technology. *www.rfidjournal.com.*
[17] Ashton, K. 2009. That 'Internet of Things' Thing. *www.rfidjournal.com.*

2.1.2 Machine to Machine (M2M)

The European Telecommunication Standards Institute (ETSI) doesn't use the word IoT but instead discusses a similar concept, M2M, to describe the process of machines operating using sensors and the Internet for connectivity:

> *"Machine-to-Machine communications is the communication between two or more entities that do not necessarily need any direct human intervention. M2M services intend to automate decision and communication processes." – ETSI*[18]

The M2M approach actively deny the users of having any role within an IoT network, instead claiming that no direct human intervention is required for most IoT.[18]

2.1.3 Users

Hewlett-Packard included the user and the context to the concept of IoT in a research study from last year, stating that the IoT refers to the unique identification and "*Internetization*" of everyday objects. This allows for human interaction and control of these "*things*" from anywhere in the world, as well as device-to-device interaction without the need for human involvement.[19]

What's noteworthy about this statement, besides an extended remark that acknowledges potential human interaction, is how the Internet enables interactions that aren't bound to specific localizations but can rather be used anywhere in the world.

Another advocate of including users in the perspective of IoT is the electronics manufacturer Philips. In one of their official blog post they claim how cloud-enabled consumer devices can produce solutions that are more meaningful and help improve people's everyday lives by not only serving people's needs, but also react to them - which in turn creates a better value for consumers.[20]

2.1.4 Computer-Supported Cooperative Work (CSCW)

Computer-Supported Cooperative Work (CSCW) has been huge part of HCI since it was introduced 30 years ago. In short, CSCW could be described as the collaborative activities and coordination that can be achieved through the means of computers.[21]

While the field of CSCW hasn't provided an explicit definition of IoT, CSCW provide an expressed vision for the emerging technology commonly known as IoT. CSCW talk of a conceptualization of the cooperation between objects and spaces as a cooperative work for the IoT that includes; coordination mechanisms, differences across contexts, common information spaces, and awareness.[22]

---

[18] European Telecommunications Standards Institute. 2010. Machine-to-Machine communications; M2M service requirements. *www.etsi.org*.
[19] Hewlett-Packard Enterprise. 2015. Internet of Things Research Study: 2015 report. *www.hp.com*.
[20] Nota, P. 2016. How the Internet of Things empowers us all. *www.philips.com*.
[21] Grudin, J. 1994. Computer-Supported Cooperative Work: History and Focus. *Computer 5: p. 19-26.*
[22] Robertson, T. & Wagner, I. 2015. CSCW and the Internet of Things. *ECSCW 2015, p. 19–23, Oslo, Norway.*

2.1.5 Proposed Definition of IoT

To help better understand how the broad concepts and many interpretations of IoT relates to this study, a specified definition of IoT will be utilized throughout this paper for pragmatic reasons. A definition that follows:

The Internet of Things consists of a network of things, each with a unique identification, that are connected to the Internet and may or may not be embedded with sensors. Characteristics of the Internet of Things is that it can be used anywhere in the world as long as the included things are all connected to the Internet. These things use machine-to-machine communication that may or may not be regulated through human interaction.

2.2 Usability

Usability is usually considered comprised of the three distinct aspects of effectiveness, efficiency, and satisfaction. They are distinct in the way they have proven to have no correlation, meaning all three aspects needs to be included during a usability testing in order for a usability test to account for the overall usability.[23]

2.2.1 ISO 9241-11

The International Organization of Standards (ISO) published a document in 1998 containing requirements describing to which extent a product can be used by a specific user to achieve a specific goal with effectiveness, efficiency and satisfaction. ISO 9241-11 explains the advantage of measuring usability in terms of user performance and satisfaction which are measured in terms of; to which degree the goal is fulfilled, the resources which have been used to fulfill the goals and to which degree the user regard the use of the product acceptable.[24]

- *Effectiveness:* Accuracy and completeness with which users achieve specified goals[24]. Example; If the goal is to turn on three lights then effectiveness is reached if all lights are turned on. If only two lights are turned on, effectiveness in the system is not reached.
- *Efficiency* Resources expended in relation to the accuracy and completeness with which users achieve goals[24].
  Example; The amount of time required to turn on the three lights.
- *Satisfaction:* Freedom from discomfort, and positive attitudes towards the use of the product[24]. Example; The user's overall attitude regarding turning on the three lights. If the users feel any kind of discomfort or enjoyment in using the system.

[23] Frøkjær, E. Hertzum, M. & Hornbæk, K. 2000. Measuring Usability: are effectiveness, efficiency and satisfaction really correlated? *CHI '00, p. 345-352, New York, USA.*
[24] International Organization of Standardization. 1998. ISO 9241-11:1998 Ergonomic requirements for office work with visual display terminals (VDTs) -- Part 11: Guidance on usability. *www.iso.org*

2.3 Visual Attention

Attention is the concentration of awareness on some phenomenon to the exclusion of other stimuli[25]. Attention is an effective way to deal with and understand a certain phenomenon, such as a task or object, while withdrawing other visual stimuli.

When talking about visual attention, it's usually described as the process in which a person moves their eyes and focus on a particular part in their visual field.

2.3.1 Eye Movement

Even though the human vision appears to be stable, the eyes are constantly moving to construct a complete image of what's in the visual field[26]. Most often, a new point of focus in our visual field will divert a person's attention, redirecting the person's focus and concentration to that new point[27].

Visual attention and a person's eye movements follow three steps;[27]

1. When presented with a stimulus, such as an image, most of the scene is seen through the peripheral vision. At this stage the scene is mostly in low resolution, though interesting features directs attention for further inspection.
2. Attention is disengaged from the foveal location and the eyes quickly reposition to the region containing the feature that attracted attention.
3. Attention is the engagement to inspect a specific feature at high resolution.

Even though the model doesn't cover all aspects of visual attention, it's a comprehensive basis to describe the cognitive aspect of visual attention.

2.3.2 The History of Eye Tracking

The history of eye tracking dates back to the late 1800s. Back then, eye tracking meant attaching a plaster of paris with mechanical linkages to recording pens directly to the participants' cornea. A method that often left the participant in discomfort.[28] This apparatus was too invasive to be widely used, though in 1901 a non-invasive eye tracking technique was introduced that used the light reflected from the eccentric surface of the cornea to track a person's eye movements[29]. As technology progressed, video recordings in eye tracking studies helped researchers answer basic hypothesis of how the brain and visual system work together[27]. It wasn't until 1947 eye tracking was used to explore how users interact with a system. One prominent study was conducted in which they tracked pilots' eye movements while the pilots were maneuvering an airplane[30]. Even with the advances made in eye tracking during the

[25] McCallum, W. 2015. Attention (Psychology). *www.britannica.com.*
[26] Bergstrom, J. Romano. Shall, Andrew. 2014. Eye tracking in User Experience Design. *Morgan Kaufmann.*
[27] Duchowski, T. Andrew. 2007. Eye Tracking Methodology. *Springer.*
[28] Young, R. Laurance. Sheena, David. 1975. Behavior Research Methods & Instrumentation: Survey of eye movement recording methods. *Springer.*
[29] Dodge, R. Cline, S. 1901. The angle velocity of eye movements. *American Psychological Association.*
[30] Fitts, M. P., Jones, E. R. & Milton, L. J. 1950. Eye movements of aircraft pilots during instrument-landing approaches. *Aeronautical Engineering Review.*

1970s, there wasn't many publications indicating eye tracking could be used to improve usability. A reason behind this might've been that a few minutes of eye tracking data could take days to analyze.[31]

2.3.3 Modern Eye Tracking

The eye tracking technology as it's known it today was introduced in the 1990s and applied in the field of HCI, a field in which it still has a strong presence. With modern, non-invasive eye trackers there is no longer a need to have a highly advanced understanding of human physiology, engineering or computer science to operate the technology.[32]

Modern eye trackers are classified into being remote or mobile.[33] Both techniques track where and for how long a user look at a specific point. Remote eye trackers are desktop versions often placed underneath a screen and tracks the user's eye movement at a distance, therefore no invasive devices need to be attached to the user. They often have a high rate of collecting eye movements, though requires the person to be seated in front of the eye tracker. It's often used in analyzing user interfaces.

Mobile eye trackers are head-mounted devices. They can be used almost anywhere and are often used for tracking tasks, allowing the user to walk and look around freely. Mobile eye trackers are not as precise as remote ones and need the user to wear a pair of glasses, thus making it a bit invasive. The gaze data from mobile eye trackers can together with recently released software be automatically mapped to a picture of the scenery, drastically cutting the time to analyze the data[34].

Commonly used visualizations are heat maps (figure 3) and gaze plots (figure 4). Areas of interest (AOI) (figure 5) are used to analyze various elements of a visual scene, with geometrical shapes indicating different elements in the user's gaze. This can be used to see in which order, how long and how often the different elements were viewed. [32]

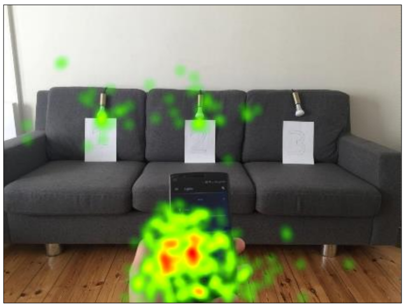

Figure 3. Heat map from one of the pilot tests

Figure 4. Gaze plot from one of the pilot tests

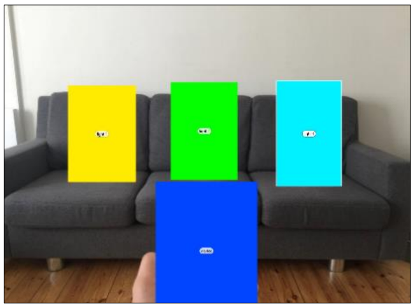

Figure 5. Defined AOIs from one of the pilot tests

[31] Jacob, R & Karn, K. 2003. Eye Tracking in Human-Computer Interaction and Usability Research: Ready to Deliver the Promises (Section Commentary). *In The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research, p. 573-603.*
[32] Bergstrom, J. & Shall, A. 2014. Eye tracking in User Experience Design. *Morgan Kaufmann. Waltham, USA.*
[33] iMotions Inc. 2016. Eye tracking - The Definitive Guide. *www.imotion.com.*
[34] Tobii AB. 2016. Tobii Pro Glasses Analyser. *www.tobii.com.*

2.4 Previous Research

Listed below are some of the studies in HCI that has touched on similar subjects found in this study. Interaction with everyday lighting systems, eye tracking the visual attention, and visual separation in multiple display systems.

2.4.1 Interaction with everyday lighting systems

How user interact with everyday lighting systems was explored in a study 2013 where Philips Hue was one of the systems tested. The aim with the study was to identify the underlying motivation for using and adapting light to come up with generic principles in designing control interfaces. The study used a set of different interaction interfaces, from cubes that upon rotation changed the lightning to a smartphone app. The result showed that using a smartphone was preferred due to the fact that users expressed that a smartphone is practically a part of their body.[35]

2.4.2 Eye Tracking the Visual Attention

In 2011 a study was conducted in which the participants were to perform a set of tasks in which they switched their visual attention between a smartphone screen and a projection on a wall or floor. A mobile eye tracker (Tobii Glasses) was used for measuring the participant's visual attention between the tasks. By using information from both the projection and the smartphone the participants had to solve a set of tasks. Some participants had both the smartphone screen and projection in their field of view while others only had the smartphone screen in their field of view. The projection was present but the participants needed to move their head in order to see it. The results showed that there was a greater change of the participants' visual attention when the projection wasn't in their field of view than when the projection was present. A result which would have been overlooked if not for the eye tracking technology.[36]

2.4.3 Multiple Displays Systems

Microsoft Research conducted a study in 2003 regarding how systems that include multiple displays often distribute information at a wider angle than single display systems. These displays are then often placed at different depths or framed by physical bezels, which then introduce physical discontinuities in the presentation of information. These systems force the user to change their visual attention to perceive all information that is at hand. The results demonstrated that when separating information between multiple displays at different depths but within the visual field there was a diminished effect on the performance at about a ten percent decrement. About half of the participants expressed higher satisfaction and preferred using the smaller display for their primary task.[37]

---

[35] Offermans, S.A.M., van Essen, H.A. & Eggen, J.H. 2014. User interaction with everyday lighting systems. *Personal and Ubiquitous Computing, Vol. 18(2014), No. 8, p. 2035-2055.*

[36] Cauchard, J. R., Löchtefeld, M., Irani, P., Schoening, J., Krüger, A., Fraser & M. & Subramanian, S. 2011. Visual seperation in mobile multi-display environments. *UIST '11, p. 451-460. ACM. New York, USA.*

[37] Desney, S. & Czerwinski, M. 2003. Effects of Visual Separation and Physical Discontinuities when Distributing Information across Multiple Displays. *INTERACT'03 M.Rauterberg et al.(Eds.) Published by IOS Press,(c) IFIP, p.252-255.*

3. Methodology

3.1 Research Design

Research design is the strategy of choosing what type of study, in terms of data collection and analysis methods, that best address the research problem and provides an answer to the research question[38]. The study type should support an operational approach aligned with the practical implications of the research question. The research question uses a setup in which a comparison must be made between two different conditions. The research question also intends to use usability as a measurement. A third aspect characterizing this thesis research question is its presumed use of IoT devices that enables novel ways to live our everyday lives. Technology that have yet to be utilized by a broader audience other than early adopters[39].

3.1.1 Research Strategies

Based on the three aspects of the research question, a lab experiment design strategy provides the means to best accommodate all aforementioned prerequisites of answering the research question.

Cluster sampling is an effective method to recruit participants for a study[38]. Variables can be measured with a pair of eye tracking glasses which record the gaze, visual field, and sound during an experiment. When collecting data about the participants' perceived satisfaction, a researcher can't simply observe the participant during an experiment. Subjective values need to be followed up by a post-test questionnaire[40].

3.1.1.1 Lab Experiment

The research design of a lab experiment is dedicated to imposing controls on variables so that the impact of a specific variable can be measured. Lab experiments are often associated with laboratories that are equipped with sophisticated tools and technology to measure variables with extreme precision. The lab experiment should be considered a strategy rather than a method as it doesn't have a specific way of describing how a study is to be conducted. The lab experiment only details what is of importance, namely the measurement and causal relationship between variables. In a lab experiment there is usually a need for a moderator that guides the participant through the different tasks. The presence of a moderator can contribute to what is known as the observer effect, which states that people are more likely to alter their behavior if they're aware of that they're being watched. This would likely be true regardless of the moderator's presence, as the very nature of an experimental environment suggests some degree of monitoring. To best avoid the observer effect is to have as little interaction with the participants as possible.[38]

[38] Denscombe, M. 2003. The Good Research Guide *(2nd ed.) Open University Press. Philadelphia, USA.*
[39] Rowland, C. Goodman, E. Charlier, M. Light & A. Lui, A. 2015. Designing Connected Products: UX for the Consumer Internet of Things. *O'Reilly Media, Inc.*
[40] Sauro, J. & Lewis, J. R. 2012. Quantifying the User Experience: Practical Statistics for User Research. *Elsevier, Inc.*

3.1.1.1.1 Variables

The purpose of an experiment is to identify the causal factors between variables by introducing or excluding factors between groups or situations. Variables are being measured and groups/situations are being observed.[41]

3.1.1.1.2 Between-Subject Design

A between-subjects design uses different groups of participants where each group will be getting a different treatment and no participant can be part of both groups. Advantages is that there's less risk of fatigue and subjects learning the test by using different subjects in each group. The disadvantage is that recruiting and running more subjects could be more time consuming and costly.[42]

3.1.1.2 Cluster Sampling

Cluster sampling is the method of focusing on a particular cluster of something that the researcher wishes to study. The idea is to use a natural occurring place while retaining the principles of random selection and the laws of probability in a cost- and time efficient way. While some clusters have a higher probability than others of including people with certain defining traits or experiences, selections made in that cluster are still chosen at random. Cluster sampling is useful when the researcher wants to study a specific population or specific location.[41]

When it comes to sample sizes in usability testing, the Nielsen Norman Group argue that studies that measures usability in quality measures (efficiency of use, user errors and subjective satisfaction) rarely need to exceed 20 users[43]. In 2010 a survey was made that asked usability testers how many users they tested in their most recent study. The results showed that a majority of the studies had between 6-11 participants while the median was 12 users[44].

3.1.1.3 Eye Tracking

Eye tracking, as a method, helps researchers understand visual attention. The technology makes it possible to detect where users look, for how long and the path their eyes follow. It can be used in numerous fields, from cognitive psychology to marketing. For user experience research, eye tracking help researchers gain a more complete view of the user experience.[42]

3.1.1.4 System Usability Scale (SUS)

The System Usability Scale (SUS) is a well-established usability questionnaire, developed as part of the usability engineering programme in integrated office systems development at Digital Equipment CO Ltd, that is often used due to being reliable and low cost. It consists of 10 questions that accounts for the subjective assessments of usability, and as such are generally used immediately after a respondent has had a chance to use the system that's being evaluated (See Appendix 1). Calculating the SUS generates a score ranging between 0-100.[45] The benefits of SUS is that it generates measurable data on a Likert-scale that can be used for statistical analysis. It's also been proven a working and free method to statistically represent the perceived usability of a system from the perspective of the user.[43]

---

[41] Denscombe, M. 2003. The Good Research Guide *(2nd ed.) Open University Press. Philadelphia, USA.*
[42] Duchowski, T. Andrew. 2007. Eye Tracking Methodology. *Springer.*
[43] Nielsen, J. 2006. Quantitative Studies: How Many Users to Test. *www.nngroup.com*
[44] Sauro, J. 2010. How Many Users Do People Actually Test? *www.measuringu.com*
[45] Brooke, J. 1996. SUS: a 'quick and dirty' usability scale. *in Usability Evaluation in Industry, p. 189-194, Taylor & Francis.. London, UK.*

While SUS is a free evaluation tool to use, the score itself requires specific knowledge to be interpreted. To someone unfamiliar with SUS the number on its own doesn't say much regarding the user's thoughts regarding the system. As a reference, a SUS score is often compared to an existing database as a point of reference to give a system a grade ranging between F to A+ to help interpret a SUS score (see table 1). The average score for the SUS is considered to be 68 points where a higher score is considered to represent a higher level of perceived usability.[46]

Tabel 1. Curved Grading Scale Interpretation of SUS Scores

SUS Score	Grade	Percentile Range
84.1 - 100	A+	*96-100*
80.8 - 84	A	*90-95*
78.9 - 80.7	A-	*85-89*
77.2 - 78.8	B+	*80-84*
74.1 - 77.1	B	*70-79*
72.6 - 74	B -	*65-69*
71.1 - 72.5	C+	*60-64*
65 - 71	C	*41-59*
62.7 - 64.9	C-	*35-40*
51.7 - 62.6	D	*15-34*
0 - 51.7	F	*0-14*

3.1.2 Alternative Research Strategies

A part of the research strategy that's as important as choosing what approach to pursue is contemplating all the available options to determine what's not worth pursuing. Many methods rely on the perceptions of the subjects that are being studied. To paraphrase the famous anthropologist Margaret Mead who summarized why subjects perceived reality is a weak metric to solely rely on for social research; what people say, what people do, and what people say they do are entirely different things.[47]

3.1.2.1 Alternative to Lab Experiment

There are two different types of experiments. Choosing between a field or lab experiment mainly comes down to the level of control required to ensure that the clausal relationship between the measurements in use are correlated and that the result is valid[44]. In the case for this study, it's also a matter of finding 'fields' to observe. As previously stated, the amount of users familiar with the Philips Hue are rather scarce. Another aspect obstructing the use of field studies are the ethical aspect of recording the experiment. Philips Hue are marketed towards a domestic consumer market. Doing a field experiment would hence require visiting peoples' homes. Seeing as recording the sessions are a necessity to accurately review the data collected, a field research would have to impose on the privacy of recording in peoples' private homes. Surveilling a person in the privacy of their home is widely considered a questionable ethical practice, regardless the level of consent[48]. A lab experiment also offers the highest

[46] Sauro, Jeff & Lewis, James. R. 2012. Quantifying the User Experience: Practical Statistics for User Research. *Elsevier Inc.*
[47] Mead, M. 1928. Coming of Age in Samoa. *William Morrow & Co.*
[48] Johannesson, P. & Perjons, E. 2014. An introduction to Design Science. *Springer.*

precision and credibility when measuring variables between groups where only one independent variable is to be changed between the experimental group and control group[35].

Surveys, along with case studies, action research and ethnography, also face the challenge of requiring actual situations to collect data from[35]. The presumed scarcity of the Philips Hue system in Swedish homes and issues of privacy makes neither type suitable in comparison to a laboratory experiment where the right conditions of people and situations can be constructed and setup.

Grounded theory is an approach that is getting more popular as it's adoptable for qualitative exploratory research. Grounded Theory is also a type that requires an open mind and inductive reasoning[35], as opposed to the deductive nature of the research question and the very existence of a hypothesis, making it less suitable for this thesis.

3.1.2.2 Alternative to System Usability Scale (SUS)

Other existing industry standard questionnaires to evaluate systems besides the SUS, and also interviews, were considered. The four leading post-study usability questionnaires are the QUIS, SUMI, PSSUQ and SUS[49]. In 2004 an extensive comparison between some of the most common usability questionnaires showed that SUS is the most reliable option compared to QUIS, CSUQ, MPRC (Microsoft Product Reaction Cards) and TSQ (Tullis & Stetson's Questionnaire), even with a small number of participants (see figure 6) [50].

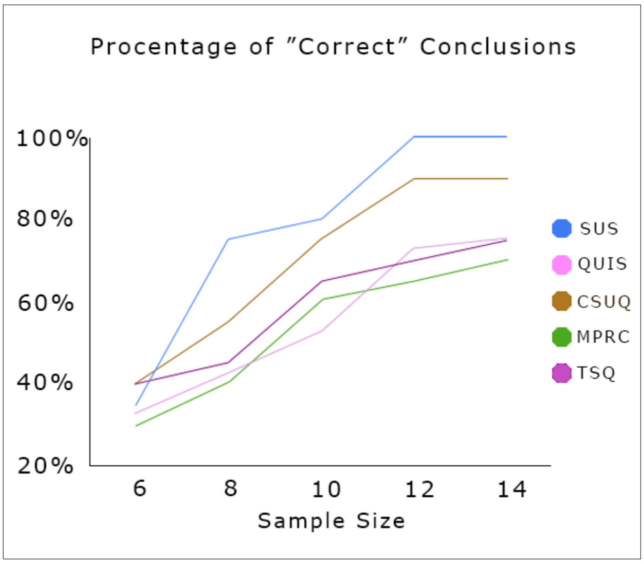

Figure 6. The percentage of sample tests that yielded the same conclusion as the analysis of the full dataet, from "A Comparison of Questionnaires for Assessing Website Usability" by T. S. Tullis and J. N. Stetson, 2004, Proceedings of UPA 2004 Conference

[49] Sauro, J. & Lewis, J. R. 2012. Quantifying the User Experience: Practical Statistics for User Research. *Elsevier Inc.*

[50] Tullis, T. S. and Stetson, J. N. 2004. A Comparison of Questionnaires for Assessing Website Usability. *Proceedings of UPA Conference '04*

3.1.2.2.1 Questionnaire for User Interaction Satisfaction (QUIS)

The Questionnaire for User Interaction Satisfaction (QUIS) was developed at the HCI Lab at the University of Maryland in 1987 and designed to guide in redesigns of systems by doing comparative evaluations of multiple UIs. The questionnaire is specifically designed for desktop software and requires a license fee to be paid in order to be used.[51] The fee and explicit focus on desktop software makes it a less viable option compared to the SUS.

3.1.2.2.2 Software Usability Measurement Inventory (SUMI)

The Software Usability Measurement Inventory (SUMI) is a questionnaire consisting of 50 questions that are available in eight different languages. The SUMI was developed by Jurek Kirakowski in 1986 and is a commercially available method mainly designed to identify problems found in an existing prototype or system. The SUMI is substantially more extensive than the SUS.[52] SUMI also requires more hands on time with the system that's being evaluated compared to SUS. This presents a significantly higher risk of the participants being unable to answer some of the questions due to not having spent enough time with the whole system. Like the QUIS, there's also a matter of economically justifying the purchase of a license fee to use SUMI.

3.1.2.2.3 Post-study System Usability Questionnaire (PSSUQ)

The Post-Study System Usability Questionnaire (PSSUQ) was developed by the Human Factors Group at IBM as their internal method of evaluating their own software. It's a 19 question questionnaire that is designed to take around 10 minutes.[53] The most common version of PSSUQ is known as the Computer System Usability Questionnaire (CSUQ)[54]. The PSSUQ requires more knowledge to be interpreted and has been scientifically proven to be less accurate in creating valid results compared to the SUS[55].

3.1.2.2.4 Interviews

Interviews can generate data that answers how the participants perceive the satisfaction of using a system post-experiment. Interviews, however, face the same issue as surveys where it relies on the perceptions of the users being interviewed and generates qualitative data that needs interpretation from the researcher to be statistically viable - reducing any chance of reproducing the test with the same reliability. Interviews also enhances what is known as the observer effect (further discussed in 3.1.3.1.2 Test Environment), which should be avoided to the greatest extent possible when designing an experiment[56].

---

[51] Chin, J. P., Diehl, V. & Norman, K. 1987. Development of an instrument measuring user satisfaction of the human-computer interface. *ACM CHI'88, 213-218. Washington, DC.*

[52] Karowski, J. and Corbett, M. 1993. SUMI: the Software Usability Measurement Inventory. *British Journal of Educational Technology. Volume 24, Issue 3, 210 – 212.*

[53] Lewis, J. 1995. IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction, 7.1, 57-78.*

[54] Sauro, Jeff & Lewis, James. R. 2012. Quantifying the User Experience: Practical Statistics for User Research. *Elsevier Inc.*

[55] Tullis, T. S. and Stetson, J. N. 2004. A Comparison of Questionnaires for Assessing Website Usability. *Proceedings of UPA Conference '04*

[56] Denscombe, M. 2003. The Good Research Guide. *(2nd ed.) Open University Press.*

3.1.2.3 Alternative to Between-Subject Design

Within-subject design differs from between-subject design by letting the same participants partake in both groups. In some studies, it is considered inappropriate as the system in evaluation might have been learnt by the participant as they use it, making experience a possible contributing factor when partaking in the test a second time under a different circumstance.[57]

3.1.2.4 Alternative to Cluster Sampling

Instead of using a probability sampling method of clusters a purposive sampling method could ease the process of finding suitable participants for a study, though a non-probability sampling method is a bigger departure from generating data that can be used to generalize the results.[55] A cluster sample provided the necessary means to both create a somewhat random sample while convenient enough to be carried out in one place during a short amount of time, which is highly suitable for lab experiments.

3.1.2.5 Alternative to Eye Tracking

An alternative to eye tracking the participants to measure the variables would be to use a think-aloud method. Two of the most common think-aloud methods are the concurrent think-aloud (CTA) and retrospective think-aloud (RTA) methods. The CTA method requires the participants to account for what they're thinking of and what they're looking at simultaneously as they're thinking or looking at something[58]. It has, however, been proven that the CTA method change how users attends to a given task which in turn may alter their gaze pattern. RTA requires a lot from the participants as they're expected to account for what they thought of or looked at after the experiment had taken place.[59] The risk of collecting the wrong measurement is considerably higher compared to eye tracking. Eye tracking generates objective data rather than a subjective assessment from the participant as the CTA and RTA does, making eye tracking more reliable.

3.1.3 Method of Application

The method of application is the way in which the methodology was practically carried out to generate the necessary data to help answer the research question. The most important aspect to account for when doing a study that aims to quantify a user experience is the concept of chance. Chance plays a role in all sample selections and need to be accounted for when comparing data. I.e. even if a majority of users in one group scores a higher score than the users from another group, that doesn't mean that that the average score for all users is higher in the first group. To determine if two means from ratio data are significantly different is just as an important part as being able to track and know what type of data to collect.[55]

3.1.3.1 The Experiment

The purpose of the experiment was to test out the hypothesis and generate relevant data to answer the research question. That is, in short, to see if usability was affected depending on if the user was by or away from the actuators while interacting with the edge device. The IoT system in use for the study was the Philips Hue system. The Philips Hue system consist of light bulbs that were wirelessly connected to a gateway that was linked to a Wi-Fi router.

---

[57] Denscombe, M. 2003. The Good Research Guide. *(2nd ed.) Open University Press.*
[58] Cooke, L. 2010. Assessing Concurrent Think-Aloud Protocol as a Usability Test Method: A Technical Communication Approach. *IEEE Transactions of professional Communication, 53.3, 202-215.*
[59] Bojko, A. (2005). Eye Tracking in User Experience Testing: How to Make the Most of It. *In Proceedings of UPA '05. Montreal, Canada.*

The edge device in use was an LG G4 H815 running Android 6.0. The official Philips Hue app[60] on the edge device made it possible for the participants to manage the actuators, no matter if the actuators were visually present or not.

In the experiment there were an experimental group and a control group. The experimental group had both the actuators and edge device present while the control group only had the edge device present. There were a set number of tasks that were identical in both groups. The tasks were designed to cover a set of functionalities present in the app. This generated data that were mapped to the concepts of efficiency and effectiveness that were measured in quantitative data and later used for analysis in the Statistical Package for the Social Science (SPSS). The participants' perceived satisfaction of the application was collected using a SUS. The SUS score were later analyzed in SPSS. The participants were equipped with a pair of Tobii Pro Glasses 2 that recorded and tracked the participant's visual distribution during the experiment. The eye tracking technology verified the visual distribution of the users.

The experimental script (appendix 2) was written in accordance to general guidelines provided by the Nielsen Norman Group for conducting eye tracking studies[61]. Each participant signed a consent form prior to participation, stating the conditions for the experiment, their role in the test, and the purpose of the study (appendix 3).

3.1.3.1.1 Variables

The experiment had one independent variable that differentiated the experimental group from the control group, and remaining four being the measured dependent variables.

Independent Variables:

- *Visual Access:* Visual access to the actuators. This was the manipulated variable that was only present to the experimental group.

Dependent variables:

- *Gaze Outside the Screen, Rate* (Ratio) Number of times a participant looks outside the edge device.
- *Task Completion, Rate* (Nominal) Measured in overall completion rate. Completed / Not completed.
- *Task Completion, Time* (Continuous) Measured and rounded in seconds.
- *SUS, Score* (Ratio) The SUS score.

In the experiment, three of the variables were directly mapped to each of the central concepts of usability.
Task Completion, Time accounted for the efficiency. *Task Completion, Rate* accounted for the effectiveness, and the *SUS, Score* accounted for the satisfaction.
Gaze Outside the Screen, Rate was a measurement used to see how the difference of having the actuators visually present or not affected the visual distribution of the user.

[60] Philips Consumer Lifestyle. 2016. Philips Hue gen 1. *Google Play Store*
[61] Nielsen, J & Pernice, K. N.d. How to Conduct Eyetracking Studies. *www.nngroup.com*

3.1.3.1.1.1 Outliers

While collecting the data, some data points usually stand out from the normality. These data points are called outliers and can be defined as observations which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism.[62] There's a great debate whether outliers should be excluded or included in the analysis of datasets. A general rule is that if a data point is ≥2 standard deviations from the mean the data point should be reviewed by the researchers to determine if there was a specific mechanism behind the data point being an outlier and if so, that data point should be omitted and the reason for omission should be stated.[63]

3.1.3.1.2 Test Environment

The test took place at the Interaction Design Lab (ID-lab) at the Department of Computer and System Sciences (DSV) at Stockholm University, Kista. As the goal of the study was to examine how the visual attention given to the actuators affected the usability of the IoT system, it was important to neglect the risk of personal experiences and preferences of the subjects affecting the results. To further ensure the actuators was the contributing factor, the ID-lab was stripped of as many objects as possible. During the test, there was a chair for the subjects to sit on and a table to rest their arms against. There was also a digital single-lens reflex camera (DSLR) mounted to a tripod located by the actuators to help capture if the participants were successful in completing the set tasks. For the experimental group, the DSLR was located behind the participant, able to catch both the interactions of the participant and the actuators. For the control group, the actuators and DSLR were located in a room next door, away from the participant. The chair and table was located in the part of the ID-lab that provided the best lighting condition and least amount of potentially distractive elements, a setup that was shared between both groups. For the experimental group, the actuators were located in a horizontal line in front of that chair and table (see figure 7). The control group was facing a blank wall, as the actuators were not present. The lighting in the room was adjusted to a degree that enabled the actuators to be seen regardless of their level of dimness, and so that the eye tracker could accurately track the participants' gaze.

The researchers were assigned the same role in every test. One acted as the moderator, responsible for managing the communications with the participant during the test. The other acted as both a recruiter and technician. As a recruiter, the responsibility was to randomly recruit participants in the vicinity. As a technician, the responsibility was to keep track of the eye-tracking recordings and DSLR footage.

[62] Hawkins, D. 1980. Identification of Outliers. *Springer Netherlands.*

[63] Albert, B. & Tullis, T. 2013. Measuring the User Experience: Collecting, Analyzing, and Presenting. *Morgan Kaufmann.*

Figure 7. The test environment for the experimental group

3.1.3.1.3 Pilot Test Implementations

A total of 13 pilot tests were conducted. The first four were made in an environment other than the final one to test out the equipment and script. The last nine both verified that the researchers knew their parts by heart and that no further details outside those given in the script had to be explained to the participants. The script was changed to further describe the purpose eye tracking had in the test, as some of the participants initially believed they would use eye tracking to control the actuators. Some participants were timider than others, so the script were changed to further outline the importance of exploring the full app. The names of the lights were changed from being positional (left, middle, right) to being numeric (1, 2, 3). The time it took to interpret left from middle compared to interpret 1 from 2 was considerable less varied in the latter.

3.1.3.1.4 Participants

The most important prerequisite for the subjects to participate was a prior experience of using a smartphone. Since the variables measured would be affected if a user couldn't operate the device on a basic level (i.e. navigational swipes and touch to press), it was important that all participants had the same basic understanding of a smartphone. Due to the application only being available in English, it was also required that the participants had a basic knowledge of the English language. In order to make sure the participants recruited weren't part of the exclusion criteria, each participant was initially subjected to a verification test, specifically made for this study[64]. The verification app was designed to make sure the participants grasped the operations of the smartphone in use.

3.1.3.1.4.1 Recruitment of participants

The sampling method chosen was in direct relation to the conditions of the test. The test environment was available for three days and the sampling size of ten participants per group is a reflection of what could be acquired within that time frame, considering the limitations of having access to the lab environment. The resources required to contact respondents and make them travel to the location of the experiment would entail considerable time from the participants. Instead, the cluster sampling method was considered best suited to generate the samples. The concept of randomness was still at play, seeing as the researchers had no control over what people would be present at the university during the time of the experiment. The drawback of randomly recruiting subjects from a site, such as DSV, is that DSV

[64] Gustafsson, F. 2016. Verify Experiences – Smartphone. *Google Play Store.*

itself has a higher probability of consisting of a certain group of people, namely students. An advantage of recruiting people from the same university with similar educations was the increased likelihood of the participants being somewhat homogenous. Considering the resources available for this study and location of the experiment, the cluster sampling method was deemed most appropriate.

3.1.3.2 The Tools

The tools used to collect and analyze the data was a combination of both software and hardware. The Tobii Pro Glasses 2 were worn by the participants to record their visual distribution during the test along with the software Tobii Pro Glasses Controller. The test recordings were encoded using the software Tobii Pro Glasses Analyzer and the data was analyzed in SPSS. The Philips Hue system was the evaluated IoT system.

3.1.3.2.1 Tobii Pro Glasses 2

The Tobii Pro Glasses 2 are designed to capture human behavior anywhere. It does this using a light weighted head unit that's equipped with four eye cameras that captures the visual surroundings (82° vertical and 52° vertical angle) and gaze of the user wearing the device at either 50 or 100 Hertz in a 1920x1080 pixel resolution with a frame per second (fps) rate of 25. The gaze tracking technique uses corneal reflection and dark pupil tracking in addition to the binocular camera. The head unit is equipped with a microphone and it's wired to a recording unit that saves the recorded data to a memory card. The head unit and recording unit are designed to be mobile and enable an unrestricted movement for the user, in terms of both head- and body movement (see figure 8).[65]

While mobility is a big selling point of what makes the Tobii Pro Glasses 2 great for behavioral research, the primary functionality that's of use in this test is rather its capability to track the gaze of a user across multiple connected devices.

Figure 8. The Tobii Pro Glasses 2 head unit and the recording unit

[65] Tobii AB. 2015. Tobii Pro Glasses 2. *www.tobii.com.*

3.1.3.2.2 Tobii Pro Glasses Controller

The controller software ran on a tablet computer (Microsoft Surface Pro 4 running Windows 10 Pro) which enabled live viewing, calibration of the glasses, recording functionalities and video export.[40] The software was used during the test to get a live view in order to calibrate the glasses for each participant, start and stop the recordings and ensure a fluid flow throughout the experiment.

3.1.3.2.3 Tobii Pro Glasses Analyzer

The Tobii Pro Glasses Analyzer is the companion software for post-analysis and visualization of the Tobii Pro Glasses 2 data[66]. The software ran on a laptop computer (ASUS UX31A-C4043H running Windows 10 Pro). The software was used to track the participants' visual attention, map the participants gaze with real world mapping to visualize the data in gaze patterns.

3.1.3.2.4 Philips Hue

The Philips Hue system consisted a gateway, three light bulbs and their smartphone app. Both the smartphone app and the light bulbs were located in the same internal network, therefore there was no need for the router to be connected to an external network. The smartphone app was used to dim the light, switch color of the light, and change the names of the light bulbs.

3.1.3.2.5 SPSS

SPSS is an extensive software for predictive analytics that was be used to evaluate, analyze, and report statistical relationships between variables in terms of correlation and significance[67]. The results was measured against one another to help answer the hypothesis. Depending on the kind of variable used, different methods of analysis was chosen.

3.1.3.2.5.1 Independent Sample t-test

The t-test is a statistical test to see if there's a significance between two sets of data. It uses the means and standard deviation of each set of data to calculate the likelihood that any difference between the two sets of data are random or not. The t-test is also well suited for this study that could be considered a small-scale research, unlike the chi square that works best with larger sample sizes.[68]

3.1.3.2.5.2 Levene's F test

The purpose of the Levene's F test is to assess the equality of variances for a variable that has been calculated for at least two groups. This was used to test the null hypothesis of a population variance as being homogeneous by calculating if the p-value was below 0.05. If it were, the null hypothesis would be rejected and it could be concluded that there was a significant variance in the groups measured. [69] The assumption of homogeneity of variance is important to measure when doing a t-test.[68]

3.1.3.2.5.3 Cohen's d

Cohen's d is an effect size that uses a comparison between two means, making it appropriate for statistical tests such as the t-test. The effect size measured how practically significant an effect was of a sample, as opposed to the statistical significance of the p-value. The effect size of Cohen's d ranges

[66] Tobii. 2015. Tobii Pro Glasses Analyzer. *www.tobii.com*
[67] IBM. n.d. IBM SPSS Software. *www.ibm.com*
[68] Sauro, Jeff & Lewis, James. R. 2012. Quantifying the User Experience: Practical Statistics for User Research. *Elsevier Inc.*
[69] Leverne, H. 1960. Robust tests for equality of variances. *In Contributions to Probability and Statistics (I. Olkin, ed.) 278–292. Stanford Univ. Press, Palo Alto, CA*

between small (0.2), medium (0.5) and large (0.8).[70] Cohen's d works as a complement to the t-test to see if the standard deviations and the sample size is similar in both groups. Cohen's d provides a benchmark for how to interpret the score, which makes it preferable to other alternatives (i.e Glass' Delta and Hedge's *g).*

3.1.3.2.5.4 Skewness and Kurtosis

Skewness is the measurement of the lack of symmetry in a distribution. A perfect normal skewness will have the value of zero, where any value on either side of zero can be considered a tail. Skewness measures the relative size of the two tails in a dataset. Kurtosis is a measurement accounting for the combined size of these two tails. Together, the skewness and kurtosis are used to find the outliers of a dataset that might otherwise skew the result.[71]

In a study made in 2010, it was recommended that a thumb of rule for the values of the skewness and kurtosis to be considered sufficiently normal for the purposes of conducting a t-test should range between <|2.0| (skewness) and <|9.0| (kurtosis) [72]. Any outliers affecting the skewness and kurtosis level beyond these values need to be examined to make sure they don't skew the whole dataset that's being analyzed. In case such an outlier exists, a researcher need to account for its removal[68].

3.1.3.2.5.5 Pearson Product-Moment Correlation Coefficient

Different from the t-test, the Pearson product-moment correlation coefficient tells how closely two variables are connected and can be used with interval and ratio data. The range of the correlation coefficient is somewhere between -1 and +1, where -1 represents a perfect negative correlation, +1 represents the opposite, and a 0 represents no correlation. It's considered a reasonable correlation if the numbers span between ± 0.3 and ± 0.7.[73]

3.1.3.3 The Analysis

The test recordings were imported to the software Tobii Pro Glasses Analyzer. Based on the recordings showing of the participants' eye movements and performances, each test had the dependent variables measured.

Task Completion, Time (Continuous):

A task was considered initiated as soon as the researcher started reading out the task, and finished once the participant claimed they felt finished by uttering "Done". The number of seconds between those points were measured.

[70] Cohen, J. 1992. A power primer. *Psychological Bulletin, 112, 155-159.*
[71] Kerlinger, F. & Lee, H. 2000. Foundations of Behavioral Research. *4:th ed. Cengage Learning. Belmont, USA.*
[72] Schmider, E. Ziegler, M. Danay, E. Beyer & L. Buehner, M. 2010. Is It Really Robust? *In Methodology European Journal of Research Methods for the Behavioral and Social Sciences 01/2010; 6(4):147-151.*
[73] Denscombe, M. 2003. The Good Research Guide *(2nd ed.) Open University Press.*

Task Completion, Rate (Nominal):

The completion rate was measured in a binary matter, where a task was considered complete if:

- Task 1a: Light bulb number 2 is turned on at the highest capacity, no matter the color.
- Task 1b: Light bulb number 2 is turned off.
- Task 2a: Light bulb number 1 shines in a red hue, the intensity is irrelevant.
- Task 2b: Light bulb number 2 shines in a green hue, the intensity is irrelevant.
- Task 2c: Light bulb number 3 shines in a blue hue, the intensity is irrelevant.
- Task 2d: All three light bulbs are turned off.
- Task 3: Light bulb number 2 has its name changed into either "Hello" or "hello".

If the condition for a task wasn't met, that task was considered incomplete.

Gaze Outside the Screen, Rate (Ratio):

The number of times the participant shifted their attention from the edge device.

SUS, Score (Ratio):

The result from the participants SUS was calculated in order with existing SUS documentation[74].

3.2 Research Ethics

There were many aspects of ethics that needed to be accounted for before, during and after carrying out the study. There are four principles described to help operationalize the traditional research principle that states; the end doesn't justify the means in the pursuit of knowledge.[75]

Principle 1: Protect the Interests of Participants.

The interest in this sense ensured that no participant came to any physical, social or mental harm as a direct or indirect result of participating in the study[75]. A precaution made in terms of protecting the interests of participants was to let every participant do the test. 5% of the population can't be tracked using eye tracking. To minimize the risk of offending participants whose eyes couldn't be tracked, even the participants with low gaze samples was allowed to continue with the test, although excluded from the final dataset.

Principle 2: Ensure That Participation is Voluntary and Based on Informed Consent.

Participation was completely voluntary throughout the conduction of the study. Each participant had to read and sign a consent form prior to participating (see appendix 3). The consent form stated all the conditions for the study and details were provided in accordance to the common practice of writing a consent form[76].

Principle 3: Operate Openly and Honestly.

This principle was achieved towards not only the participants, but also toward the scientific community that the results presented were true, transparent, and had a high credibility[76]. The participants were given as much information about the study as possible without risking skewing the results. They were informed

---

[74] Brooke, J. 1996. SUS - A quick and dirty usability scale. *www.hell.meiert.org/core/pdf/sus.pdf*
[75] Johannesson, P. & Perjons, E. 2014. *An introduction to Design Science*. Springer.
[76] Johannesson, P. & Perjons, E. 2014. *An introduction to Design Science*. Springer.

that the test measured usability. They were informed that the eye tracking technology was used to track their eye movement. They weren't informed as to how usability would be measured in terms of time lapsed, completion rate and the SUS score given. Informing them could've potentially altered the way in which they interacted with the IoT system.

Principle 4: Comply with Laws.

This one is rather self-explanatory. A researcher is never above the law[57]. Always act accordingly.

The aspect of ethics towards the scientific community is usually achieved through validity and reliability, both of which are further detailed below.

3.2.1 Validity

To minimize the risk of prior experiences affecting the results, each participant was subjected to a verification test making sure they all had the required skills necessary to use the app. The verification app was specifically created for this study to include all gestures and features found in the tasks used in the experiment[77].

Translations for non-native English speakers has been proven to maintain the same level of validity, even in conjunction with the original SUS[78]. During the pilot tests the specific word "cumbersome" was misunderstood by two participants. When a participant asked about the word "cumbersome", it was replaced with the word "awkward", as suggested by the founder of SUS himself[79].

3.2.2 Reliability

Reliability refers to the consistency of producing data that generates equal results on different occasions[80]. In this study, reliability was foremost favored to generalizability when designing the test environment. Many of the downsides recognized of doing a field experiment instead of a lab experiment was accounted for in the heading 3.1.2.1 Alternative to Lab Experiment. Additionally, the field experiment would have had many variables that would remain unaccounted for and that probably might have differed between the subjects. On the plus side of a field study, however, findings of a field study are more applicable outside the framework of the research compared to a lab experiment, giving it a higher level of generalizability. By making sure there are as many constant variables as possible in the experiment, the validity also increases.[80]

[77] Gustafsson, F. 2016. Verify Experiences – Smartphone. *Google Play Store.*
[78] Finstad, K. 2006. The System Usability Scale and non-native English speakers. *Journal of Usability Studies.*
[79] Brooke, J. 2013. SUS: A Retrospective. *Journal of Usability Studies.*
[80] Denscombe, M. 2003. The Good Research Guide *(2nd ed.) Open University Press.*

4. Results

4.1 Visual Distribution

The experimental group (N = 10) was associated with a gaze outside the screen rate value of M = 11.5 (SD = 4.70). By comparison, the control group (N = 10) was associated with a numerically lower rate of M = .1 (SD = .32). To test the hypothesis that users devotes less visual attention to the smartphone app if the light bulbs are present, an independent sample t-test was performed between the experimental group and control group, measuring the variable *"Gaze Outside the Screen, Rate"*. The distributions in the experimental group were sufficiently normal for the purposes of conducting a t-test[81], with skewness of .467 and kurtosis of - 4.777. For the control group the variable *Gaze Outside the Screen* was constant, leaving it omitted in terms of skewness and kurtosis. Additionally, the assumptions of homogeneity of variances was tested and satisfied via Levene's F test, F (18) = 17.94, p = .000. The independent samples t-test was associated with a statistically significant effect t (18) = -7.66, p = .000. Thus, the experimental group were associated with a statistically significant lack of visual attention given to the smartphone app compared to the control group.

Cohen's d was estimated at 3.42, which is well beyond the large effect based on Cohen's guidelines[82]. A graphical representation of the statistical distribution is displayed in figure 9.

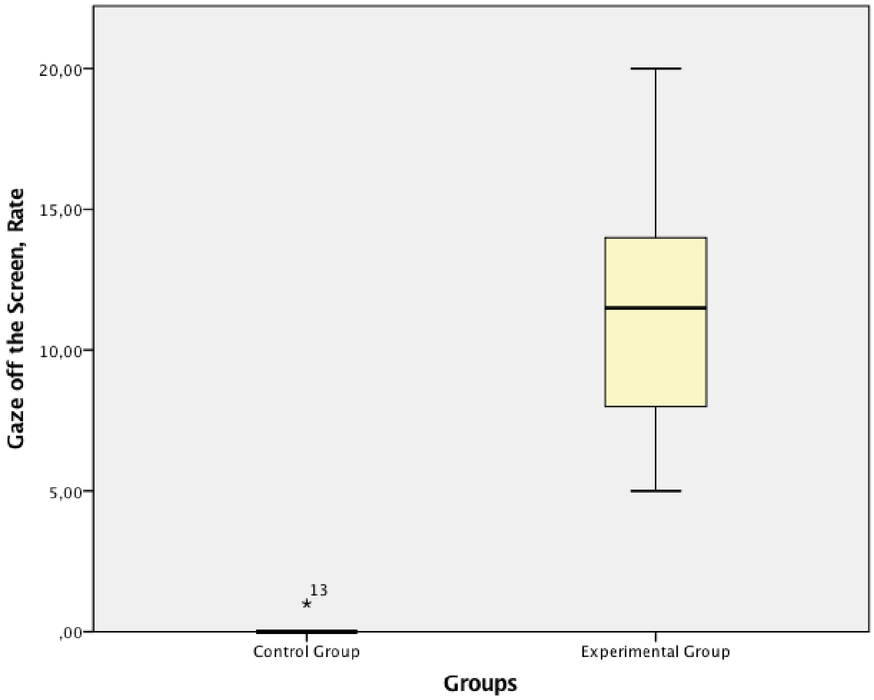

Figure 9. Box Plot Diagram for Gaze Outside the Screen, Rate variable between the experimental group and control group

[81] Schmider, E. Ziegler, M. Danay, E. Beyer & L. Buehner, M. 2010. Is It Really Robust? *In Methodology European Journal of Research Methods for the Behavioral and Social Sciences 01/2010; 6(4):147-151.*
[82] Cohen, J. 1992. A power primer. *Psychological Bulletin, 112, 155-159.*

The gaze plots, of both task 1a and 1b combined, clearly illustrate the difference in visual distribution between the two groups. The participants chosen to exemplify this was chosen based on their scores being closest to the median scores in each group respectively. The numbers show the chronological order in which the participants shift their attention to a new fixation point, while the size of the circle illustrates the duration of the fixation point being in focus. Larger circle equals a longer fixation time. As can be seen in figure 10, the visual attention is more or less equally distributed between the smartphone app and the light bulb number two in the experimental group. In the control group, all visual attention was given to the smartphone app as can be seen figure 11.

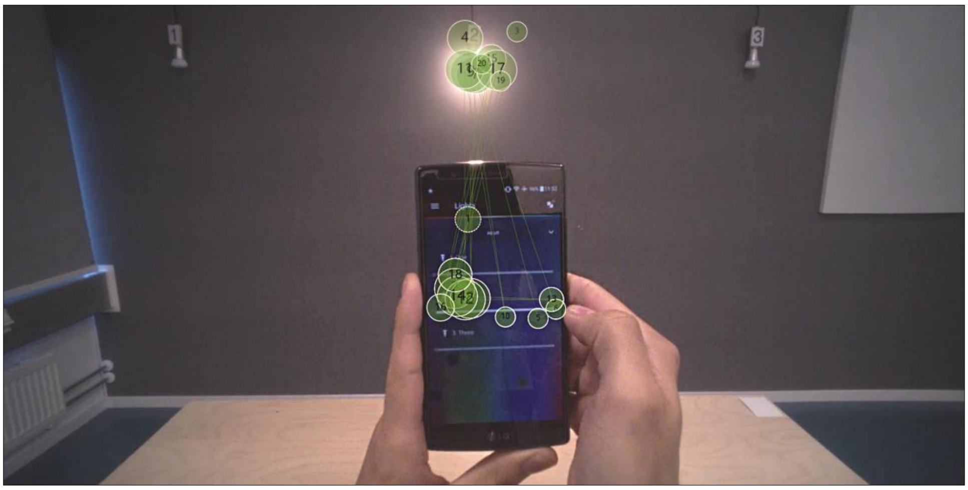

Figure 10. Gaze Plot of the first task. Participant from the experimental group

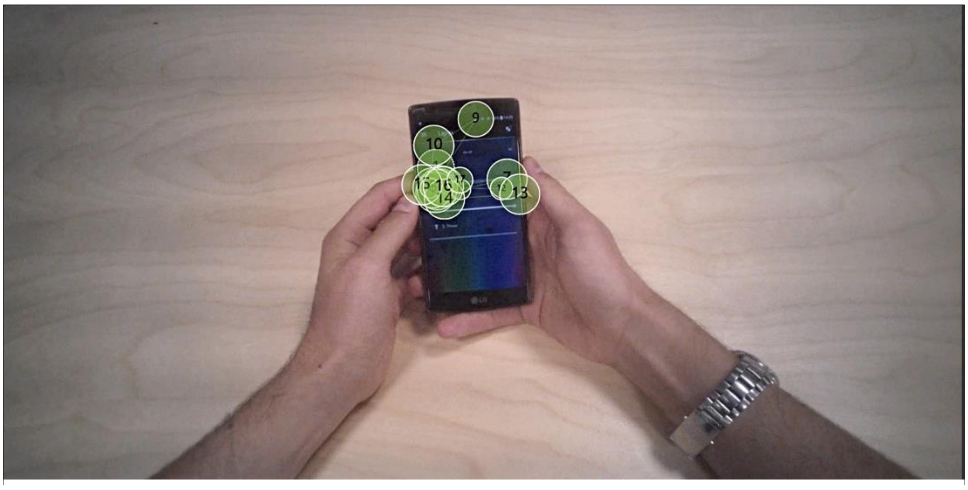

Figure 11. Gaze Plot of the first task. Participant from the control group

4.2 Effectiveness

The experimental group (N = 10) was associated with an effectiveness value of M = 6.60 (SD = .70). By comparison, the control group (N = 10) was associated with a numerically higher rate of M = 6.80 (SD = .42). To test the part of the hypothesis that claims that a visual confirmation of the light bulbs enhances effectiveness, an independent sample t-test was performed between the experimental group and control group, measuring the variable *"Task Completion, Rate"*. The distributions in both the experimental and control group were sufficiently normal for the purposes of conducting a t-test[83]. The experimental group had a skewness of -1.658 and kurtosis of 2.045. The control group had a skewness of -1.779 and kurtosis of 1.406. Additionally, the assumptions of homogeneity of variances was tested but could not be satisfied via Levene's F test, F (18) = 2,82, p = .111. The independent samples t-test was not associated with a statistically significant effect t (18) = -.775, p = .449. Thus, the experimental group could not be associated with a statistically significant enhanced level of effectiveness compared to the control group.

Cohen's d was estimated at .35, which is slightly above the small effect based on Cohen's guidelines[84]. A graphical representation of the statistical distribution is displayed in figure 12.

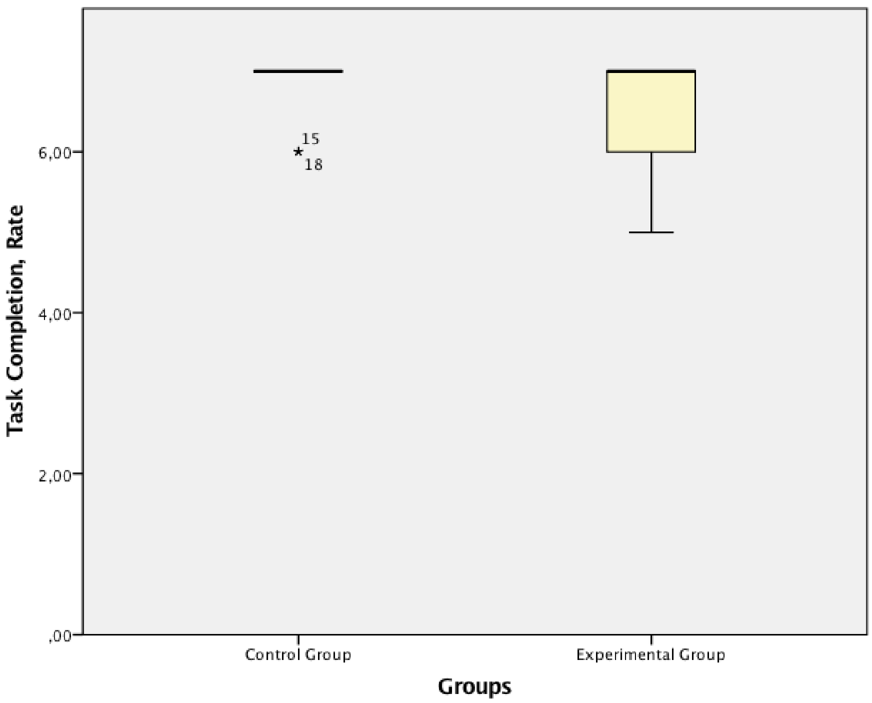

Figure 12. Box Plot Diagram for the "Task Completion, Rate" variable between the experimental group and control group

[83] Schmider, E. Ziegler, M. Danay, E. Beyer & L. Buehner, M. 2010. Is It Really Robust? *In Methodology European Journal of Research Methods for the Behavioral and Social Sciences 01/2010; 6(4):147-151.*
[84] Cohen, J. 1992. A power primer. *Psychological Bulletin, 112, 155-159.*

4.3 Efficiency

In the control group one participant's time for completion was greater than three standard deviations from the mean. Reviewing the recordings indicated that the participant failed to understand that he/she were to utter "done" upon completion of the task 1a. This was realized long after the task actually was completed. As a result, the participant was removed completely from the analysis of efficiency.

The experimental group (N = 10) was associated with an efficiency value of M = 155.40 (SD = 49.41). By comparison, the control group (N = 9) was associated with a numerically lower rate of M = 141.56 (SD = 48.31). To test the part of the hypothesis that claims that a visual confirmation of the light bulbs enhances efficiency, an independent sample t-test was performed between the experimental group and control group, measuring the variable *"Task Completion, Time"*. The distributions in both the experimental and control group were sufficiently normal for the purposes of conducting a t-test[85]. The experimental group had a skewness of .361 and kurtosis of -.839. The control group had a skewness of .720 and kurtosis of -.352. Additionally, the assumptions of homogeneity of variances was tested but not satisfied via Levene's F test, F (17) = .032, p = .861. The independent samples t-test was not associated with a statistically significant effect t (17) = - .616, p = .546. Thus, the experimental group could not be associated with a statistically significant enhanced level of efficiency compared to the control group.

Cohen's d was estimated at .28, which is slightly beneath the small effect based on Cohen's guidelines[86]. A graphical representation of the statistical distribution is displayed in figure 13.

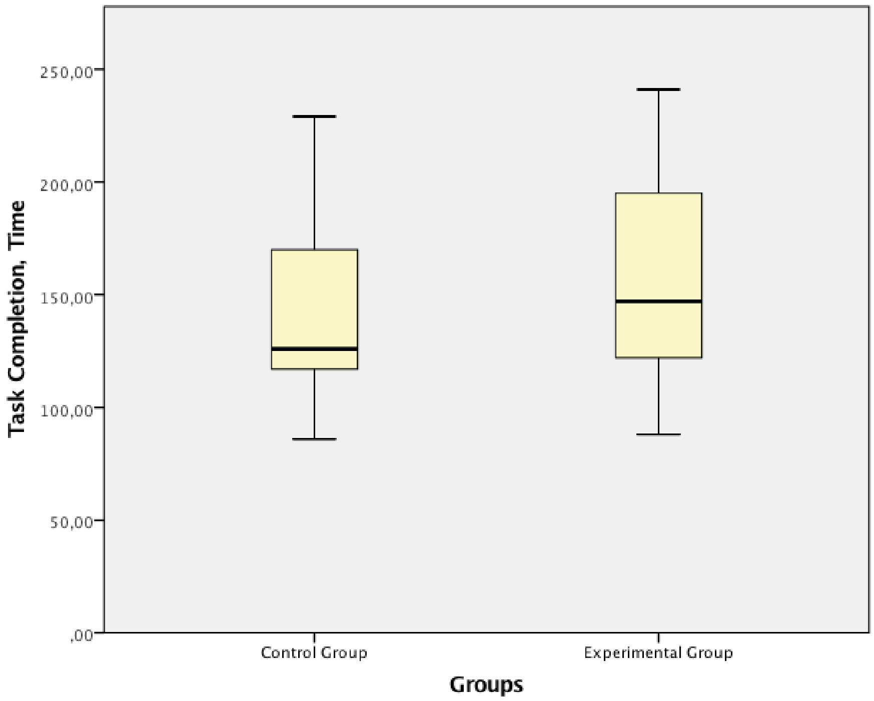

Figure 13. Box Plot Diagram for the "Task Completion, Time" variable between the experimental group and control group

[85] Schmider, E. Ziegler, M. Danay, E. Beyer & L. Buehner, M. 2010. Is It Really Robust? *In Methodology European Journal of Research Methods for the Behavioral and Social Sciences 01/2010; 6(4):147-151.*
[86] Cohen, J. 1992. A power primer. Psychological Bulletin, 112, 155-159.

4.4 Satisfaction

In the control group one participant's SUS score was beyond two standard deviations from the mean. After completing the test, the participant expressed his/her thoughts and ideas that revealed that the participant was evaluating the system as a developer rather than an end-user. As a result, the participant was removed completely from the analysis of satisfaction.

The experimental group (N = 10) was associated with a SUS score of M = 63.25 (SD = 15.81). By comparison, the control group (N = 9) was associated with a numerically higher rate of M = 74.72 (SD = 9.80). To test the part of the hypothesis that claims that a visual confirmation of the light bulbs enhances satisfaction, an independent sample *t*-test was performed between the experimental group and control group, measuring the variable *"SUS, Score"*. The distributions in both the experimental and control group were sufficiently normal for the purposes of conducting a *t*-test[87]. The experimental group had a skewness of .018 and kurtosis of -2.063. The control group had a skewness of -.790 and kurtosis of -.914. Additionally, the assumptions of homogeneity of variances was tested and satisfied via Levene's F test, F (17) = 7.68, p = .013. The independent samples *t*-test was close, but not quite associated with the statistically significant effect of .05, but rather t (17) = 1.874, p = .078. Thus, the experimental group were close to being associated with a statistically significant decreased level of satisfaction compared to the control group.

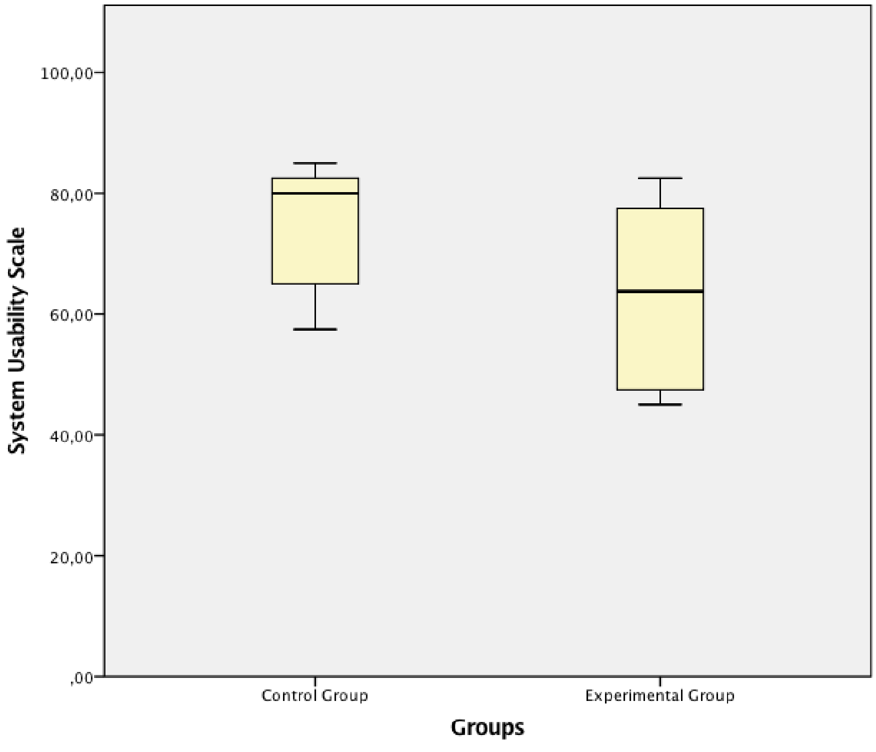

Figure 14. Box Plot Diagram for the "SUS, Score" variable between the experimental group and control group

[87] Schmider, E. Ziegler, M. Danay, E. Beyer & L. Buehner, M. 2010. Is It Really Robust? *In Methodology European Journal of Research Methods for the Behavioral and Social Sciences 01/2010; 6(4):147-151.*

Cohen's *d* was estimated at .87, which is above the large effect based on Cohen's guidelines[88]. A graphical representation of the statistical distribution is displayed in figure 14.

To verify if there was a relationship between the *SUS, Score* and *Gaze Outside the Screen*, a Pearson product-moment correlation coefficient was computed. There was a negative correlation between the *SUS, Score* and the *Gaze Outside the Screen* variables, a significant value couldn't be showed, r = -,321, N = 20, p = ,168. Pearson's r shows that there's a relationship between a lower SUS score and Gaze Outside the Screen but not the cause. The effect and the significance of the finding can't be statistically proven. The grade given by each of the participants can be seen in table 2.

Table 2. The SUS scores converted to grades in each group

Experimental Group	Grade	Percentile Range	Control Group	Grade	Percentile Range
82,5	A	90-95	85	A+	96-100
82,5	A	90-95	82,5	A	90-95
77,5	B+	80-84	82,5	A	90-95
75	B+	80-85	80	A-	85-89
70	C	41-59	80	A-	85-89
57,5	D	15-34	75	B	70-79
50	F	0-14	65	C	41-59
47,5	F	0-14	65	C	41-59
45	F	0-14	57,5	D	15-34
45	F	0-14			

[88] Cohen, J. 1992. A power primer. *Psychological Bulletin, 112, 155-159.*

5. Discussion

5.1 Conclusion

Research Question: In what ways do the usability of an edge device differ if the user gets visual feedback from the actuator when interacting with the edge device?

Hypothesis: People who use a smartphone app to interact with a light bulb devotes less visual attention to the smartphone app in favor to the light bulb if present, which enhances the usability of the IoT system in terms of effectiveness, efficiency and satisfaction.

While the results confirmed the first part of the hypothesis, regarding the visual distribution being different depending on the presence of the actuators, the dataset was unable to disprove the null hypothesis which states that a change in visual attention does not enhance the overall usability. Rather, the *t*-variables showed that it was the control group, as opposed to the experimental group, that indicated a greater value of both efficiency and satisfaction while the effectiveness was equally distributed between both groups. Neither have a significant *p*-value, though the value for satisfaction is relatively close where $p = .078$.

To answer the research question in full, one must break it down. Usability consists of effectiveness, efficiency and satisfaction. All three of which showed different levels of affectedness depending on the visual feedback given from the actuators.

Both groups showed an equal numerical high level of effectiveness. Out of the seven tasks the experimental group and control group scored a mean of 6.6 and 6.8 respectively, meaning that most participants in both groups were able to finish every task. The few occasions where a task wasn't completed could be argued to be due as a result of the participant either misinterpreting the task or simply being more prone to get irritated and give up. The fact that each task involved the light bulbs giving some sort of visual feedback as to how they were being manipulated seemed to have no significant impact on the effectiveness of the IoT system as a whole.

Efficiency was measured in the time it took to finish each task, which was slightly different between the groups. One participant from the control group was removed due to his/her score deviating too many points from the mean. Upon inspection of the recording of the removed session, it's evident that the participant initially failed to understand that he/she had to utter "done" after completing a task. After the first task, however, the participant performed closer to the mean in regards to the efficiency. The experimental group timed a mean of finishing all tasks at 155.4 seconds, while the control group scored a smaller mean with the amount of time being 141.6 seconds, implying that the control group were more efficient at executing the tasks judging by their mean. Important to discuss here is the high *p*-value of .546 and the *SD* being close to 50 in each group. While surely a result of the number of participants being too few to make out a pattern in any of the groups, it could also point towards unaccounted variables such as specific personality traits or learnability being a contributing factor. Seeing as the difference between both groups are rather low, it could also mean that the difference reflects the added amount of time the experimental group spent looking at the added visual feedback from the light bulbs before determining if they felt finished with a task or not. In other words, distributing visual attention to the visual feedback given from the light bulbs reduced the efficiency.

The satisfaction was measured in the post-test SUS questionnaire. After converting the answers from a Likert scale to ratio data, the means in the two groups were computed. The average score for the SUS questionnaire is considered to be 68 points where a higher value is considered to be a higher level of perceived satisfaction. In contrast to the hypothesis; the satisfaction was higher in the control group, which was associated with a mean score of 74.72 compared to the control group's score of 63.25. The mean score of the control group's perceived satisfaction was higher than approximately 70% of all products tested in the SUS database while the experimental group's score was approximately higher than 30%[89]. None in the experimental group gave the Philips Hue system the highest grade while four participants gave it the lowest grade, resulting in the mean grade being a *C-*. In comparison to the control group where one participant gave the Philip Hue system the highest grade while none gave it the lowest grade, resulting in the mean grade being a *B*. It need to be taken in account for that p = .078, proving that the results aren't statistically significant, though relatively close. According to the Pearson product-moment correlation coefficient there's a relationship between how often a participant looked away from the smartphone app and the SUS score. With the Pearson's r = -.321 and Cohen's d = .87, indicates that not having visual feedback is correlated and has a large effect on being more satisfied.

In conclusion, there were no significant differences in effectiveness and efficiency depending on if the user was given visual feedback by the light bulbs or not when interacting with the smartphone app. The satisfaction, however, was shown to decrease when interacting with the smartphone app if there was a visual feedback given by the light bulbs and the user had to distribute their visual attention across multiple devices.

5.2 Limitations

Every research project has limitations that often needs to be addressed in regard to a given time frame of meeting a specific deadline or accommodating an economical budget. The number of participants that was recruited for this study was a direct result of both those factors, and judging by the p-values in two out of the four variables measured, more participants were required to give the results greater generalizability. Given more time to access the facilities to conduct more tests would've theoretically been preferable. Though, a larger sample size would've practically demanded more time required to analyze the collected data. Given the time frame of a bachelor thesis the chosen number of 20 participants was, however, deemed sufficient in terms of establishing extensibility for future studies to build upon. The participants were only familiar with the Philips Hue system for a short time. If the participants would have had experience of using the Philips Hue system in their own homes the results might have differed. It would also measure a situation that's more closely related to the way in which the Philips Hue system is intended to be used in an everyday setting.

The fact that all participants was unfamiliar with the system also set limitations for what could be measured. Learnability, and the pace in which people learn to understand new things differs immensely from person to person. Due to time constraints, learnability wasn't accounted for in the variable of "Time Completion, Time" which showed the largest percentile differences in standard deviations. This could be related to the concept of learnability. By having a sample population of experienced Philips Hue users, the concept of learnability could've been minimized. Recruiting participants with prior knowledge was simply not possible given the time frame of this project.

[89] Sauro, Jeff & Lewis, James. R. 2012. *Quantifying the User Experience: Practical Statistics for User Research*. Elsevier Inc.

5.3 Ethical delibirations

The experiment was designed in regard to avoid creating any emotional, social or physical impact on the participants. In terms of reproducibility, the test was created and documented in great detail to encourage a full reenactment of the test. The script can be followed through and through and the environment should be easy to recompose. A total of 9 pilot tests was conducted before deciding the consistency, and hence reliability, of the experiment was at a desired level. In retrospective, no measure of enhancing the reliability can be thought of.

In terms of validity, only one variable was deemed questionable in regards as to the results it yielded. The small number of participants leave this up for discussion but, as it stands, the question if "Task Completion, Time" is solely a reflection of the collocation of the actuators is up for debate. It could just as easily be related to the concept of learnability, one of which isn't included in this study. Hence, validity could be improved by increasing sample size and to some extent account for the concept of learnability when deciding the sample population.

For researchers, it's important to understand the concept of reflexivity and that they're part of the social world in which they aim to objectively observe and how relationship affects the ways in which they perceive phenomenon's. While the hypothesis, research question and following discussion was formulated in regards to prior experiences, perceived values and norms of the researchers, all methods was chosen to produce objective data. This resulted in a dataset that can be reviewed by anyone, regardless of social or scientific background, and interpreted as they seem fit. This also creates a greater extensibility for the result to be of use for future studies.

As mentioned in the limitations, the number of participants is the most prominent limitation which also has an effect on generalizability. The probability sampling method provides the mean to generalize the results across a larger population, yet for this study, the results would benefit from a larger sample size in order to do that. The study is also rather local, in terms of it only having examined the Philips Hue system rather than multiple IoT systems to explore the visual distribution across connected devices. While broad claims can't be made, the results provided has a comprehensive transferability for similar research that aims to explore the concept of visual distribution across devices.

5.4 Previous Research

The hypothesis was in direct contrast to the findings of the study made by Microsoft Research in 2003, where they presented a performance decrement of 10% when using multiple displays and preference of focusing on a single display. What was thought to separate this study from the one made back then is that the Philips Hue light bulbs only adds an additional layer of feedback to let users know what interactions have been made within the app, and that the growth in devices used by people for the past 10 years would have made most people more accustom to the idea of using multiple devices at once. This study and the Microsoft Research study showed a similar result where the visual feedback is preferred as focused coming from a single device rather than distributed across multiple devices. [90]

[90] Desney, S. T. and Czerwinski, M. 2003. Effects of Visual Separation and Physical Discontinuities when Distributing Information across Multiple Displays. INTERACT'03 M.Rauterberg et al.(Eds.) Published by IOS Press,(c) IFIP, 2003,pp.252-255

5.5 Current Contributions

The aim of this study was to help contribute to some of the aspects that need further study. In recap of what was requested by the HCI journal:

- Studies of novel collocated interaction settings (including 'everyday' settings and multi-device ecologies);
- Design and study of interaction techniques and systems supporting collocated interaction (including cross-device interaction, spatial interaction, mixed-presence, and groupware systems);
- Discussion of methodologies and theoretical approaches to study and design collocated interaction.

The way in which the Philips Hue enables users to interact with lights using IoT can definitely be considered novel, yet the interaction takes place in an everyday setting. Studying the relationship between how different scenarios affects the usability of this multi-device ecology has indicated how the visual attention is distributed across devices. The method in use, eye tracking, was also proven an efficient tool to help generate accurate data for this type of study.

5.6 Future Research

For future studies, it would be of interest to use the approach of research through design by using the findings made in this study to generate designs that take account for the user visually perceiving the object they manipulate. Are there designs that could help improve the usability of a system by taking advantage of the fact that the user distributes their visual attention across all connected devices?

It would be interesting to see a study that research the concept of visual distribution in Virtual Reality (VR) to see if the results found in this study also applies to VR. While the interaction setting of a VR is to a single device it enables the user to shift their visual attention by physically moving their head, reinforcing the concept of actively shifting one's visual attention to experience the entirety of a system.

Another aspect that could help contribute to the subject would be a similar test using a field experiment rather than lab experiment. The field test would require a lot more resources to pull off. Recruiting participants that are familiar with the Philips Hue system would be beneficial in two regards. Firstly, the result would have a higher validity as a result of the data being measured at the same place as where it's actually being used in its everyday setting. Secondly, only testing experienced users would minimize the risk of learnability affecting the results. The challenge of doing a field experiment would primarily be finding the 'right' sample and have a size big enough to get the right *p*-values.

Acknowledgements

We'd like to express our gratitude to our contact Jesper Högström at Tobii AB for providing us with the eye tracking technology that made this study possible. We'd also like to thank the ID-lab administrator Jinyi Wang at DSV for granting us full access to the facility where the experiment took place, and our mentor Preben Hansen for his support throughout the process of this project.

References

Albert, Bill. & Tullis, Tom. 2013. Measuring the User Experience: Collecting, Analyzing, and Presenting. *Morgan Kaufmann. Waltham, USA.*

Ashton, Kevin. 2009. That 'Internet of Things' Thing. *www.rfidjournal.com http://www.rfidjournal.com/articles/view?4986 (Accessed 2016-05-04).*

Bergstrom, J. Romano. Shall, Andrew. 2014. Eye tracking in User Experience Design. *1st ed. Waltham: Morgan Kaufmann.*

Berkers, Frank. Koers, Wietske. Colucci, Katia. Kadlec, Oskar. Puiu, Dan. Roelands, Marc. Menoret, Stephane. 2013. Vision of the future business ecosystem, new roles and models of acceptance. *www.iot-icore.eu. http://www.ioticore.eu/attachments/article/89/20130930%20iCore_D1.3_Vision_of_the_future_business_ecosystem_v1.00-1.pdf (Accessed 2016-05-04)*

Bojko, A. (2005). Eye Tracking in User Experience Testing: How to Make the Most of It. *In Proceedings of UPA '05. Montreal, Canada. https://www.researchgate.net/publication/266161907_Eye_Tracking_in_User_Experience_Testing_How_to_Make_the_Most_of_It (Accessed 2016-05-17)*

Brooke, J. 1996. SUS - A quick and dirty usability scale. *Redhatch Consulting Ltd. http://www.usabilitynet.org/trump/documents/Suschapt.doc (Accessed 2016-05-09)*

Brooke, J. 2013. SUS: A Retrospective. *Journal of Usability Studies. (8)2, 29-40. http://uxpajournal.org/wp-content/uploads/pdf/JUS_Brooke_February_2013.pdf (Accessed 2016-05-09)*

Chin, J. P., Diehl, V. & Norman, K. 1987. Development of an instrument measuring user satisfaction of the human-computer interface. *ACM CHI'88, 213-218. Washington, DC. http://www.cs.umd.edu/local-cgi-bin/hcil/rr.pl?number=87-11 (Accessed 2016-05-17)*

Cohen, Jacob. (1992). A power primer. *Psychological Bulletin, 112, 155-159. http://psych.colorado.edu/~willcutt/pdfs/Cohen_1992.pdf (Accessed 2016-05-17)*

Commodore. 1986. The X-10 Powerhouse does everything but put out the cat. *www.commodore.ca http://www.commodore.ca/gallery/adverts_other/x10_compute_jan86.jpg (Accessed 2016-05-04)*

Cooke, L. 2010. Assessing Concurrent Think-Aloud Protocol as a Usability Test Method: A Technical Communication Approach. *IEEE Transactions of professional Communication, 53.3, 202-215. http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=5556472&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D5556472 (Accessed 2016-05-17)*

Denscombe, Martyn. 2003. The Good Research Guide. *(2nd ed.) Open University Press, Philadelphia.*

Dodge, Raymond. Cline, S. Thomas. 1901. The angle velocity of eye movements. *American Psychological Association. Psychological Review. Volume 8. Issue 2. pp 145-157.*

Duchowski, T. Andrew. 2007. Eye Tracking Methodology. *(2nd ed.) London: Springer*.

Ericsson AB. 2015. Ericsson Mobility Report: Internet of Things to overtake mobile phones by 2018. http://www.ericsson.com. https://www.ericsson.com/news/2016987 (Accessed 2016-06-01)

European Telecommunications Standards Institute. 2010. Machine-to-Machine communications (M2M); M2M service requirements. *www.etsi.org. http://www.etsi.org/deliver/etsi_ts/102600_102699/102689/01.01.01_60/ts_102689v010101p.pdf (Accessed 2016-05-04)*

Evans, Dave. 2011. The Internet of Things: How the Next Evolution of the Internet Is Changing Everything. *Cisco. http://www.cisco.com/c/dam/en_us/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf (Accessed 2016-05-04)*

Finstad, K. 2006. The System Usability Scale and non-native English speakers. *Journal of Usability Studies (1)4, 185-188. http://uxpajournal.org/the-system-usability-scale-and-non-native-english-speakers/ (Accessed 2016-05-09)*

Fisher, J. 2016. HCI Journal Special Issue on Collocated Interaction: New Challenges in ‘same time, same place’ Research. *CSCW 16´ San Francisco, USA.*

Fitts, M. Paul. Jones, E. Richard. Milton, L. John. 1950. Eye movements of aircraft pilots during instrument-landing approaches. *Aeronautical Engineering Review. Volume 9. Issue 2. pp 24-29.*

Frøkjær, Erik. Hertzum, Morten. & Hornbæk, Kasper. 2000. Measuring Usability: are effectiveness, efficiency and satisfaction really correlated? *In Proceedings of the SIGCHI conference on Human Factors in Computing Systems, CHI ‘00, pages 345-352.*

Gates, Bill. 1999. Business @ the speed of thought. *1st. Essex: Pearson Education Limited.*

Grudin, J. 1994. Computer-Supported Cooperative Work: History and Focus. *Computer 5: 19-26. http://dl.acm.org/citation.cfm?id=187669 (Accessed 2016-06-01)*

Gustafsson, Filip. 2016. Verify Experiences – Smartphone. Google Play Store. *https://play.google.com/store/apps/details?id=filip.usertesting (Accessed 2016-05-09)*

Hawkins, D. William. 1980. Identification of Outliers. *Springer Netherlands.*

Hewlett-Packard Enterprise. 2015. Internet of Things Research Study: 2015 report. *www.hp.com. http://www8.hp.com/h20195/V2/GetPDF.aspx/4AA5-4759ENW.pdf (Accessed 2016-05-04)*

IBM. n.d. IBM SPSS Software. *www.ibm.com. http://www.ibm.com/analytics/us/en/technology/spss/ (Accessed 2016-05-09)* iMotions Inc. 2016. Eye tracking - The Definitive Guide. *www.imotion.com https://imotions.com/blog/eye-tracking/ (Accessed 2016-05-06)*

International Organization of Standardization. 1998. ISO 9241-11:1998 Ergonomic requirements for office work with visual display terminals (VDTs) -- Part 11: Guidance on usability. *www.iso.org http://www.iso.org/iso/catalogue_detail.htm?csnumber=16883 (Accessed 2016-05-05)*

Jacob, J.K. Robert. Karn, S. Keith. 2003. Eye Tracking in Human-Computer Interaction and Usability Research: Ready to Deliver the Promises (Section Commentary). *In The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research. 1st ed. Holland: Elsevier.*

Johannesson, Paul. & Perjons, Erik. 2014. An introduction to Design Science. *Springer. http://link.springer.com/book/10.1007%2F978-3-319-10632-8 (Accessed 2016-05-12, specific access required)*

Juniper Research Ltd. 2015. 'Internet of Things' Connected Devices to Almost Triple to Over 38 Billion Units by 2020. *http://www.juniperresearch.com. http://www.juniperresearch.com/press/press-releases/iot-connected-devices-to-triple-to-38-bn-by-2020 (Accessed 2016-05-04)*

Kanter, Theo. n.d. Smart Service For Future Internet. *DSV Research, Stockholm University. http://dsv.su.se/en/research/news/kanter-1.113279 (Accessed 2016-05-16)*

Karowski, J. and Corbett, M. 1993. SUMI: the Software Usability Measurement Inventory. *British Journal of Educational Technology. Volume 24, Issue 3, 210 – 212. http://onlinelibrary.wiley.com/doi/10.1111/j.1467-8535.1993.tb00076.x/abstract (Accessed 2016-05-17)*

Kerlinger, F. & Lee, H. 2000. Foundations of Behavioral Research. *(4:th ed.) Cengage Learning. Belmont, USA.*

Leverne, Howard. (1960). Robust testes for equality of variances. *In Contributions to Probability and Statistics (I. Olkin, ed.) 278–292. Stanford Univ. Press, Palo Alto, CA*

Lewis, J. 1995. IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction, 7.1, 57-78. http://dl.acm.org/citation.cfm?id=204774 (Accessed 2016-05-17)*

McCallum, W. Cheyne. 2015. Attention (Psychology). *www.britannica.com. http://global.britannica.com/topic/attention (Accessed 2016-05-05)*

Mead, Margaret. 1928. Coming of Age in Samoa. *William Morrow & Co, USA.*

Minerva, Roberto. Biru, Abyi. Rotondi, Domenico. 2015. Towards a definition of the Internet of Things (IoT). *IInstitute of Electrical and Electronics Engineers (IEEE). http://iot.ieee.org/images/files/pdf/IEEE_IoT_Towards_Definition_Internet_of_Things_Revision1_27MAY15.pdf (Accessed 2016-05-04)*

Nielsen, J & Pernice, K. N.d. How to Conduct Eyetracking Studies. *Nielsen Norman Group. https://www.nngroup.com/reports/how-to-conduct-eyetracking-studies/ (Accessed 2016-05-17)*

Nielsen, J. 2006. Quantitative Studies: How Many Users to Test. *www.nngroup.com https://www.nngroup.com/articles/quantitative-studies-how-many-users/ (accessed 2016-05-17)*

Nota, Peter. 2016. How the Internet of Things empowers us all. *www.philips.com http://www.philips.com/a-w/innovationmatters/blog/how-the-Internet-of-things-empowers-us-all.html. (Accessed 2016-05-05)*

Page, Larry. 2014. Where's Google going next? *TED. https://www.ted.com/talks/larry_page_where_s_google_going_next? (Accessed 2016-05-04)*

Philips Consumer Lifestyle. 2016. Philips Hue gen 1. *Google Play Store. https://play.google.com/store/apps/details?id=com.philips.lighting.hue&hl=en (Accessed 2016-06-04)*

Roberti, Mark. 2005. The History of RFID Technology. *www.rfidjournal.com http://www.rfidjournal.com/articles/view?1338 (Accessed 2016-05-04)*

Robertson, T. & Wagner, I. 2015. CSCW and the Internet of Things. *ECSCW 2015: Proceedings of the 14th European Conference on Computer Supported Cooperative Work, 19–23, Oslo, Norway. http://www.ecscw.org/2015/chp:10.1007/978-3-319-20499-4_15.pdf (Accessed 2016-06-01)*

Rowland, Claire. Goodman, Elizabeth. Charlier, Martin. Light, Ann. Lui, Alfred. 2015. Designing Connected Products. *O'Reilly Media, Inc.*

S.A.M. Offermans H.A. van Essen, J.H. Eggen. 2014. User interaction with everyday lighting systems. *Personal and Ubiquitous Computing, Vol. 18(2014), No. 8, p. 2035-2055. https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-1:v1:en (Accessed 2016-06-01)*

Sauro, J. 2010. How Many Users Do People Actually Test? *www.measuringu.com http://www.measuringu.com/blog/actual-users.php (accessed 2016-06-01)*

Sauro, Jeff & Lewis, James. R. 2012. Quantifying the User Experience: Practical Statistics for User Research. *Elsevier Inc.*

Schmider, Emanuel. Ziegler, Matthias. Danay, Erik. Beyer, Luzi. Buehner, Markus. 2010. Is It Really Robust? *In Methodology European Journal of Research Methods for the Behavioral and Social Sciences 01/2010; 6(4):147-151. https://www.researchgate.net/publication/232449663_Is_It_Really_Robust (Accessed 2016-05-17)*

Tobii AB. 2016. Tobii Pro Glasses Analyzer. *www.tobii.com. http://www.tobiipro.com/product-listing/tobii_pro_glasses_analyzer/ (Accessed 2016-05-05).*

Tullis, T. S. and Stetson, J. N. 2004. A Comparison of Questionnaires for Assessing Website Usability. *Proceedings of UPA Conference '04. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.396.3677&rep=rep1&type=pdf (Accessed 2016-05-25)*

Wolf, Gary. 1996. Steve Jobs: The Next Insanely Great Thing. *Wired. http://www.wired.com/1996/02/jobs-2/ (Accessed 2016-05-04)*

Young, R. Laurance. Sheena, David. 1975. Survey of eye movement recording methods. *Behavior Research Methods & Instrumentation. Volume 7, Issue 5. 1975. pp 397-429.*

Appendix 1 – System Usability Scale

	Strongly agree				Strongly disagree
1. I think that I would like to use this system frequently	1	2	3	4	5
2. I found the system unnecessarily complex	1	2	3	4	5
3. I thought the system was easy to use	1	2	3	4	5
4. I think that I would need the support of a technical person to be able to use this system	1	2	3	4	5
5. I found the various functions in this system were well integrated	1	2	3	4	5
6. I thought there was too much inconsistency in this system	1	2	3	4	5
7. I would imagine that most people would learn to use this system very quickly	1	2	3	4	5
8. I found the system very cumbersome to use	1	2	3	4	5
9. I felt very confident using the system	1	2	3	4	5
10. I needed to learn a lot of things before I could get going with this system	1	2	3	4	5

Appendix 2 – Experiment Manual

Hi and welcome.

In this test, we're evaluating the Philips Hue System. The Philips Hue is a consumer product developed by Philips that allows users to control their lights at home. The user can dim the lights or change the color of the lights, among other things. All of this is done using a smartphone app, and both the smartphone and lights are connected to the same network, making it possible to control the lights using the smartphone no matter if you're home or not.

The purpose of this experiment is to test the usability of the Philips Hue app. What that implies is that we're not testing you and that it's not a competition. The full system is being evaluated, so don't hesitate to explore the app in order to finish the tasks.

The test is completely optional, so if you for any reason feel like quitting you are free to do so at any time. You are wearing a pair of eye-tracking glasses that will record what you see and your eye movement. This will generate eye-tracking data for us to use later in the study. We will put the glasses on and off for you. They are rather fragile, so if you feel something is wrong or itching please tell us and we will help you remove the glasses.

The experiment has several tasks for you to complete. Each task will be presented one at a time. When you feel done with a task, simply say "done" or "finished" so I can proceed and give you another one. All tasks will be done via the app. *The entire session is designed to take around 10 minutes.*

**only to control group:* This test simulates a situation where you're away from your Philips Hue light bulbs, yet want to control them using your app. *

Please avoid from asking us any questions during the tests unless deemed necessary. If it isn't suitable to be answered, we will remain silent.

Now, do you have any question of your own before we begin?

We will calibrate the glasses before we begin with the test.
The last part before we begin the test is an important one. The glass indicates the area in which your eyes can be tracked, so it's very important that you don't look outside the frames of the glass.

Great, then let us begin with the test and remember to say "done" or "finished" whenever you feel done with a task.

Task 1a: Turn <u>on</u> the second light bulb at maximum capacity.
Task 1b: Turn <u>off</u> the second light bulb.
Task 2a: Make the number 1 light bulb Red.
Task 2b: Make the number 2 light bulb Green.
Task 2c: Make the number 3 light bulb Blue.
Task 2d: Turn <u>off</u> all three lights.
Task 3: Change the name for the number 2 light, from "Two" to "Hello".

Lastly we would like you to answer a set of questions regarding the system (hand them the questionnaire with the SUS-questions).

Thank you for participating, do you have any last comment about the study?

Appendix 3 – Informed Consent Form

Please read this consent agreement carefully before you decide to participate in the study

The purpose of this study is to evaluate the usability of smartphone controlled light bulbs (Philips Hue System).
I will contribute to the study by performing a set of tasks covering different functions of the Philips Hue application. I agree to the session and my performance being recorded using audio and visual equipment during the test.

I understand that the test is designed to take between 10-15 minutes, and that my participation is completely voluntary. I have the right to withdraw from the study at any moment simply by notifying a researcher that I wish to discontinue. Should I wish to do so prior to publication, any recorded material of my participation will be destroyed as a result of my disengagement.

I understand the data will be treated anonymously and handled confidentially by the research team.

I will receive no payment for participating in the study, as I understand that the benefit of participating in this study is that the results may help contribute to the research field of HCI.

If I have any questions or feedback regarding the study, I may contact:
Filip Gustafsson
*Tel: ****
*Email: ***@***.****

Gustav Larsson
*Tel: ****
*Email: ***@***.****

Bachelor Students
Department of Computer and System Sciences, DSV
Stockholm University

I agree to participate in the research study described above.

____
Name

____
Signature

____
Date

Stockholm University
SE-106 91 Stockholm
Phone: 08 – 16 20 00
www.su.se

Stockholm University
SE-106 91 Stockholm
Phone: 08 – 16 20 00
www.su.se

Stockholms universitet