TÉCNICO LISBOA

Tangible interaction on different technological setups

João Miguel Viana Amaral Craveiro

Thesis to obtain the Master of Science Degree in

Computer Science and Engineering

Supervisors: Prof. Ana Maria Severino de Almeida e Paiva

Prof. Francisco João Duarte Cordeiro Correia dos Santos

Examination Committee

Chairperson: Prof. Nuno João Neves Mamede

Supervisor: Prof. Francisco João Duarte Cordeiro Correia dos Santos

Member of the Committee: Prof. Mário Rui Fonseca dos Santos Gomes

November 2014

Abstract

This study investigates the differences and similarities between two tangible user interface setups used to teach logistics, the TinkerLamp and the TinkerWeb, being the latter an effort to create an affordable option to support a business. Results show that the Lamp performs consistently better among all interaction and cognitive tests but most of the times with very small and significant margins. The study also provides evidence that the TinkerWeb requires more adaptability but users clearly improve their performances. Implications of the study are discussed in terms of the business perspective, the benefits of this approach for education and future complimentary research on this topic.

Keywords

Tangible User Interfaces, Education, Logistics, Technologic setups, User experiments

Table of Contents

Table of figures

Table of tables

Table of abreviations

TUI: Tangible User Interface

GUI: Graphical User Interface

EPFL: École Polytechnique Fédérale de Lausanne

MIT : Massachussets Institute of Technology

CHILI: Computer-Human Interaction in Learning and Instruction

HCI: Human Computer Interaction

1. Introduction

The integration of our digital information systems with everyday physical objects can be achieved through technologies that enable the input of interaction dynamics and augmented reality. These are called tangible user interfaces (TUI), a concept born in the mid/late 90's, and that can have a myriad of applications (Ishii & Ullmer, 1997). Subject of a growing number of studies and implementations, TUIs already represent a strong alternative to regular UIs in very relevant contexts. The advantage of graspable and tangible interfaces relies on the idea that they enable an enactive mode of reasoning as well as empirical abstractions of sensori-motor schemes (Schneider, Jermann, Zufferey, & Dillenbourg, 2011). According to (Zufferey, 2010) the main impacts are on exploration, collaboration and playfulness of the task. Other studies raise attention to the 3D influence on perception besides usability. While the learning field is where most of these systems have been employed, domestic appliances and museums installations are also representative.

Nevertheless the development of such interfaces is known as a complex task. Capturing input from physical objects and abstracting the information to make it relevant for the system can be demanding and its combination with augmented reality feedback, which needs to be very precise in details such as calibration and feedback time, step up this process to a whole new level (Klemmer, Li, Lin, & Landay, 2004). This may be the main reason why only recently we started to see the spread of these technologies into general interest meaningful areas.

TUIs are highly dependent on technologic setups and its development and under such a fast paced environment as we have today it's important to point out the principles of design (Shaer & Hornecker, 2010) that enclose the scope we focus on:

- Tangibility and materiality
- Physical embodiment of data
- Bodily interaction
- Embeddedness in real spaces and contexts

We must also consider that although we have definitions and principles of design that help us clear out what these technologies are, the influences from other areas such as arts in general, product design or industrial design are essential. Considerations on the physical forms, used materials and other relevant features take the usual developers out of their comfort zone, create the need for more broad skills but ultimately can impact in many different important dimensions of interactions. Significant inputs also come from commercially successful areas such as entertainment as big players step into all these concepts more and more every day.

2. State of the Art

Since the dawn of tangibles interfaces in the 90's innumerous research labs and companies have expressed their interest in these technologies that interconnect the physical and digital world. We came a long way since the first wooden blocks used by Fitzmaurice to manipulate digital objects *(Fitzmaurice, Ishii, & Buxton, 1995)*.

Every year more and more systems are engineered and revealed to the public, either as research artifacts or commercial products. Open-source platforms associated with budget hardware have brought this technology to a broader public of developers and enthusiasts. These studies and products include perspectives from psychology, cognitive sciences, computer science, sociology, philosophy, and other disciplines that guide the process of design, building and evaluation of the interfaces.

The Tangible Media Group at the Massachussets Institute of Technology (MIT) is seen as the leading entity in the field with awarded projects in areas spanning from music and design to urban planning and cooking. Remarkable projects include Tangible Geospace, a map of the MIT campus projected on a table (Ishii & Ullmer, 1997). Repositioning objects such as the buildings would lead to a new self-arrangement of the map. Other tangibles enabled functions such as rotation or zooming.

The Swiss Federal Institute of Technology in Lausanne has several educational and collaborative work supporting technology using tangible interfaces. At the CHILI lab projector-camera systems are paired up with powerful recognition and simulation software to provide a compelling system to teach logistics, statics and 3D visualization. Studies on the educational dimensions of these technologies are also conducted and provide insight on what to consider and pursue while designing such systems (Dillenbourg & Evans, 2011).

The scientific community increased awareness on the subject has led to a faster development of the field. In 2007 the first conference on the topic, TEI (Tangible, Embedded and Embodied Interaction), was held at Baton Rouge. An increasing number of papers and research groups from all over the world, with interest in the area, is a reality. Actual numbers are different to obtain since this is a very multi-disciplinary area one can find projects that relate to the field coming from diverse research centers. A good example is the Reactable, an electronic musical instrument with a tabletop TUI (Jordà, 2010), that came from the Music Technology Group at Universitat Pompeu Fabra in Barcelona.

But this is not a merely academic topic. The widely known Lego Mindstorms, a kit with software and hardware to create customizable robots started as an MIT Media Lab project (Resnick, 1993) and ended up as a very successful commercial product. PixelSense, commercialized by Microsoft, is a platform that can run Windows software with the extra feature of recognizing tangibles by their foot print. Topobo is another well known tangible system that uses blocks resembling LEGO pieces that can be tracked and interpreted in many ways (Raffle, Parkes, & Ishii, 2004). Jive is an interesting platform that uses tangibles to make interaction easier for elderly users.

Open source solutions have also stepped in this field. Frameworks such as reacTIVision help developers to surmount one of the biggest obstacles in TUIs by tracking markers attached to physical objects and doing multi-touch finger tracking (Kaltenbrunner & Bencina, 2007). With different solutions and frameworks making the way from the software side, from an economic point of view, hardware is still an issue in TUI development.

Hardware is a critical part of these technologies. Projector-camera systems are the typical setup and while the projectors have remained a bottleneck economically, cameras have experienced a significant evolution. Regular cameras nowadays have the necessary features to support most systems and more advanced ones, like Kinetic from Microsoft or Creative 3D camera, that allow for image-based 3D reconstruction and gesture recognition.

3. Study Context

The Swiss educational system has a track that consists on vocational training after students turn sixteen. The so called apprentices face concrete tasks daily, where they interact with the social and physical world, but must also attend classes at school for the more theoretical concepts. An identified problem of this method is the gap experienced between what is learned in the classroom and how it applies to practical work.

Since the possibility of a broader use of TUIs has become a reality, several educational systems have been developed with the objective of addressing the perceived problem. The Computer-Human Interaction in Learning and Instruction (CHILI) lab at École Polytechnique Fédérale de Lausanne (EPFL) has been developing technologies for vocational training, namely on the field of logistics learning. The objective is to enable the integration of theoretical concepts in concrete experience. This is accomplished by using, for the example of logistics, small-scale models of a warehouse and tangible shelves as a basis for problem-solving exercises. The first system was developed for an augmented reality setup and different studies confirmed the potential of this technology (Schneider et al., 2011).

To bring this prototype from the lab to the classrooms a startup named Simpliquity was founded. Making a business out of this technology is not easy and the logistics platform was chosen as a pilot. Trying to sell these equipments to schools, with the price tag attached to the necessary hardware, was a major problem. Simpliquity worked on this side of the project and proposed a new technological setup by removing the need for a projector and improving the image recognition to work with cheaper cameras that have less resolution.

Affordability is no longer a problem but it's unclear if the results from previous experiments using the most expensive setup remain valid. Thus, the main goal of this study is to replicate some simple proceedings both in the old a new technological setup and to evaluate subjects' performance. Exact replication is a difficult task but very important as we know that factor such as representation location and speed dynamics can have a major influence in the outcome (Price, Falcão, Sheridan, & Roussos,

2009). The results will help to understand what has changed with this new approach. It will also help to identify issues that designers, researchers and developers should tackle, in order to maintain the successful features identified in the past while pursuing ubiquity of the technology.

4. Technology

As stated before, this study consists on assessing the interactions using different setups for the logistics specific learning context. Besides all the necessary equipment for the tangible interface itself, a technological support for the measuring of the dependent variables of the study is also needed. In this section we will present the main artifacts used to conduct the experience.

4.1. TinkerLamp

The TinkerLamp is a tabletop learning environment which allows apprentices to build small-scale models of a warehouse using physical objects like plastic shelves and docks. The system is made of table covered with whiteboard material and a gallows carrying a camera, a projector and a mirror. The purpose of the camera is to track the position of objects on the table and transfer this information to a computer running a logistics simulation. The projector is used to project information on the table and on top of the objects, indicating for example the accessibility of the content of each shelf or security zones around obstacles (Zufferey, Jermann, & Dillenbourg, 2008).

4.2. TinkerWeb

Recently new alternatives to the TinkerLamp have been investigated, mostly due to its expensive cost, unaffordable to most education institutions. A new solution called the TinkerWeb has been developed and introduces some major changes in the technological setup and in the interaction paradigms from before. The projection capabilities have been removed as this new setup recurs solely to a webcam to track the tagged tangibles. The setup is now simpler than ever as all it requires is the webcam and a stand. The cam points down to a surface, preferably uniform and of light color, and provided sufficient environment lightning everything is set. All the interaction information is no longer displayed as augmented reality but presented on the computer screen. The necessary software runs on a browser thanks to HTML5, which enables the access to the webcam information and other important features.

4.3. Chilli tags

Chilli tags is a cross-platform software library for the detection and identification of 2D fiducial markers developed in house. Using a camera, the position of the tagged objects can be acquired by a computer and used to virtually display information on them. These can also be used to keep track of objects on the eye gaze processing. We tagged all the shelves as well as the computer screen four corners.

Picture 1: Chili Tags

4.4. SMI mobile eye tracker

This eye tracker is presented in the form of glasses that have no lenses. The glasses frame contains cameras below the eyes that are responsible for capturing the eye movement. This is possible because besides the cameras the glasses frame also contains some leds that project imperceptible light dots on the subject's eye. The triangulation of the iris recognition and the points of light is the key. The information is easily deployed into a computer using an USB port. The ergonomics of this artifact allow capturing this sensible information with a minimal interference on the actual processes going on which is very important.

Picture 2: SMI Mobile eye tracker

4.5. Logging

An in-house framework was used to log all the relevant events during the experiments. The system allows us to output system variables from detected external events, simulation workflow or even rendering properties. All this information is stored in text files that accurately portray all the dimensions of the procedure and can be easily parsed and submitted to statistical analysis.

5. TUIs in Education

This study focuses on educational technologies that use a TUI. The field itself was born with educational concerns thus it's with no surprise that technologies for this purpose are developed and studied extensively across different dimensions (Marshall, 2007). Marshall's framework is a great corner stone to define which of these dimensions can play a significant role in this specific case.

Carefully designed tangibles can provide external representations that relate to knowledge, structure, rules, constraints and relations embedded in physical configurations (Jiajie Zhang, 1997). Appropriate representations can reduce the cognitive effort required by grouping information in the objects themselves (Larkin & Simon, 1987). The shelf models used with the TinkerLamp and the TinkerWeb

provide an external representation that better bridges the system with a real warehouse and gives a sense of proportion.

A classic result on this is a better performance achieved in the Hanoi Tower problem by the subjects using tangible representations of the towers (Jiaje Zhang & Norman, 1994).

Beyond simple isomorphisms from actual objects, multiple external representations can also have a very positive impact on performing complex tasks. In this case we clearly have an abstract representation, with all the graphics and numeric information, and a concrete representation with the graspable shelves. This is very important because novices usually can only understand more concrete representations while experts are able to operate with abstract ones. The presence and interconnection of both in the Tinker platforms enables students to better articulate abstractions and understand more advanced concepts (Ainsworth, 1999).

Collaboration, exploration, cognition & meta-cognition and playfulness are other important factors on advocating for these technologies even if they're not an object of study in this experiment. We knew from early studies the positive impacts with children (Gabrielli, Harris, Rogers, Scaife, & Smith, 2001) but more recently these have been assessed as key factors on the learning experience (Dillenbourg & Evans, 2011) and some impacts have been identified experimentally (Schneider et al., 2011). This is the kind of positive impact that we're hoping to find in the new technological setups and eventually study if the most basic aspects of interaction and cognitive processes remain valid.

6. Research Questions

The developments in the more mature and researched technological setups bring along some significant changes to the dynamics of the user-tangibles interaction. While it may seem that we're in the presence of some minor changes, some of these can have significant impacts on the users' interactions and subsequently on the outcomes of the learning processes.

Major overall experience outcomes can result from the change of augmented reality for the representation on screen and the limited resources for webcam information real-time processing. Thus we question: can the TinkerWeb be operated as effectively as the already accredited TinkerLamp?

An important definition and at this moment and from now on is performance. This evaluation will be based on speed and accuracy to hit a target or a desired position.

We will also try to understand the interaction performance upon continued usage and to go beyond simple Human Computer Interaction (HCI) by presenting the users a cognitive task. A building activity about a basic understanding of warehouse properties and continued manipulation of the tangibles towards a specific goal will be conducted.

Therefore, some other questions we pose shall be: Can the users develop strategies to better manipulate the interfaces, in particular the TinkerWeb? Are there significant performing differences on warehouse building tasks? What type of learning curves can we find for both interfaces?

We expect to find different learning curves among the different setups and also different interaction patterns. While the TinkerLamp should be easier to interact with in the first place it's possible that after some training the TinkerWeb can step up to similar interaction performances at least. Other possible constraint facts worth pointing out are the usage of the computer screen raising split attention and the commute times involved but maybe the users can find ways to overcome these, with or even without the help of the video background.

7. Method

7.1. Conditions

The different technological setups are main concerns and therefore the modalities of the experiment. One can state them as the TinkerLamp and the TinkerWeb.

This experiment considers a within subjects design. While aware of the ordering effect and how in this specific case it will probably influence the performance on the latter experimented setup, the feedback on having the experience on both equipments as well as the adaptation from one to the other are valuable information and we still have an unbiased measuring from the first interaction of each subject. Besides that there is the chance of getting stronger results while needing a smaller pool of participants. Within subject design is also great to reduce variance associated with individual differences but in this case the strong ordering effect suppresses it.

TinkerLamp	**TinkerWeb**
G1	G2
G2	G1

Table 1: Experiment conditions

Each of these groups has a total number of 5 participants so in the end there are 10 subjects tested.

The data collection from the subjects will be done before, during, in-between and after the experiment. While not interacting, a small questionnaire is conducted in order to help determine the perceptions or the experience of the subjects so far. During the experiments the dependent variables will be, as mentioned before, speed and accuracy on task completion (all measured and logged automatically by the system). A measured process variable will be the eye gaze.

We will also take into consideration some factors that might bias the outcome of the experiment. The controlled variables will be the habilitations, all fresh-man students from EPFL, and the subjects' strong hand. While the first control will be done during selection process for the experiment the second will be controlled by setting the starting shelves on the top corner of the strong hand side of the grid. Because we will be measuring speed and accuracy this may turn out to be a relevant factor.

7.2. Eye gaze capture with mobile eye-trackers

Mobile eye-trackers can be a very helpful tool to this kind of experiments for several reasons. While it won't add up to our performance measurement it will turn out essential to answer some of the sub-questions presented which can help us derive causal relations to our main measurements. The mobile feature is also very important because we ensure a minimal sight and movement obstruction during the interactions.

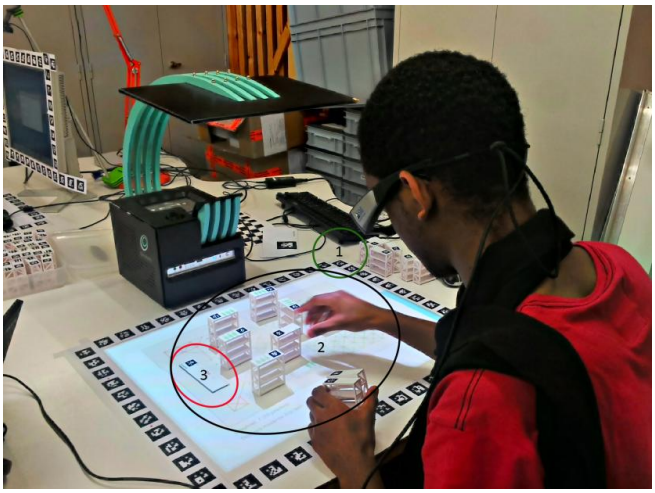
To capture the eye gaze we placed ChilliTags on each of the shelves to be used during experiment and also on the four corners of the computer screen. This way we may later analyze the data to better understand the dynamics during the interaction and how the subjects' attention spreads across the tangible interface. It will be important to understand any developed strategies by the subjects to deal with the split attention issue when using the TinkerWeb.

Eye-tracking results will be crucial in the part of this study concerning a split attention effect. As presented before in the technology section, the TinkerWeb splits it's representations between the screen and the table. The big tags around these two areas allow for the logging system to automatically detect which part the subject is focused on.

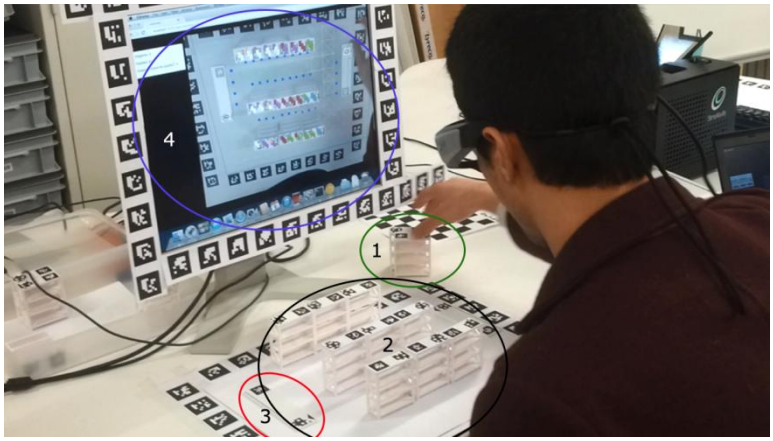
7.3. Tasks

The tasks to be performed during the experiment were designed to be as similar as possible on both the TinkerLamp and the TinkerWeb. The development was first made to the web platform which presents more restrictions and then translated to the Lamp framework. The tasks are sequenced so that first the users get a grip on the basic interaction, then on a more extended situation that introduces some warehouse specific features and finally a task that combines it all and represents a full experience of the platform.

The following images depict the two setups used for the experiments and are included here with the objective of illustrating in context some concepts I'll be using during experimental analysis. 1: Shelves starting point; 2: Interaction area; 3: In-Dock (the in- and out-dock are placed symmetrically on opposed sides of the interaction area); 4: Web screen interface. The big printed tags are used with the eye-tracking videos to measure how much time a user spends looking at each part of the interface.



Picture 3: TinkerLamp setup



Picture 4: TinkerWeb setup

7.3.1. HCI Tasks

The first set of tasks consists on moving one or more shelves towards a presented target. The target looks like a shadow of the shelf which should be moved straight the top of it. The system is continuously detecting shelf movement and when the shelf is close enough to the target (a small error margin was contemplated) the trial is considered complete. The trials will be divided into 4 trial types with increasing complexity.

First trial type (1 shelf):

- 4 translations
- 4 translations with 90º rotation
- 4 translations
- 4 translations with 90 degree rotation

Second trial type (2 shelves):

- 5 translations with 90 degree rotation

Third trial type (5 shelves):

- 5 translations with 90 degree rotation
- 5 translations with 90 degree rotation and alignments

Fourth trial type:

- 2 translations with 90 degree rotation and alignments

The workflow process of loading tasks and trials is automatic but the shelves must be replaced to start position manually at each trial.

Picture 5: Moving a shelf to a target on the TinkerLamp

7.3.1.1. Fitts' Law

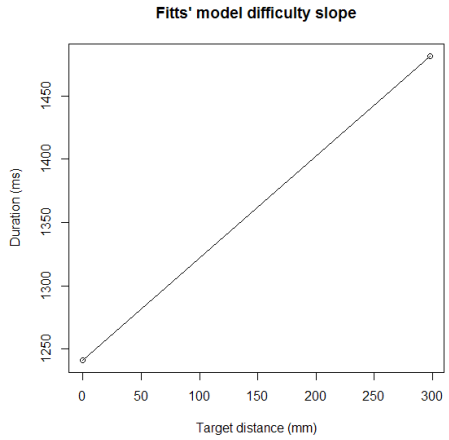
Reference values are important to guide us through the process with some anchors on what to expect and compare against. For reference values on HCI tasks one can use Fitts' Law, frequently used in HCI and ergonomics studies, to estimate the necessary time to complete each trial (MacKenzie, 1992). Having these reference values one can make comparisons with the experiment data to look out for external factors influencing our experiment if systematic deviations arise.

$$(1)\quad T = a + b\log_2\left(\frac{D}{W} + 1\right)$$

It's important to notice that this formulation of the law is used for 1 dimension analysis only. I recurred to the improvements made by MacKenzie and Buxton by using their SMALLER-OF model and the Shannon formulation (MacKenzie & Buxton, 1992). This proposed model is quite simple and seems to perform well at the level of the most complex tasks tested on their paper. The idea is to consider the error for the smaller of the 2 dimensions of a target for the W parameter. In this case it will be the height precision variable which I defined in the software to be 5mm.

The 'a' and 'b' parameters can be calculated using experiment results. These are simply the parameters of a linear regression on the data. To obtain 'b', which can be seen as the difficulty slope

for our different trials, one can consider the average time for the easiest and for the hardest trial (out of the first 16 only, trial type 1, to avoid multiple targets unnecessary complexity). The targets considered were the easiest possible one with translation only, the closest one to the shelves starting point (Trial 6) and the most difficult one with translation plus rotation, opposite corner of the detection area from the shelves starting point (Trial 12). A linear model that opposed the average times obtained by subjects on these specific trials against the distance between the two trials targets was used. Because this distance needs to also account for rotation, using the dimensions of one shelf and basic trigonometry one shall obtain a rotation and then sum it to the distance. The results for the TinkerLamp and the TinkerWeb were quite close an average value was considered. Parameter 'a' can easily be obtained from the model after getting 'b'.



Picture 6: Difficulty slope for the HCI tasks

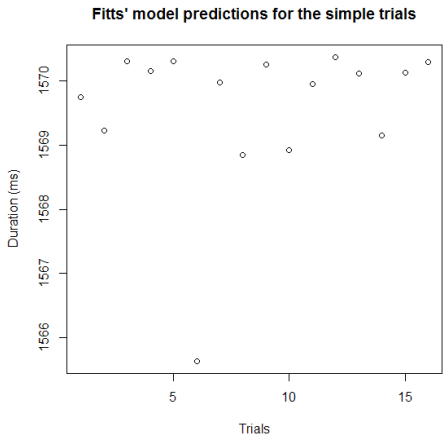
	Intercept ('a')	Slope ('b')
Linear model coefficients	1559.6500	0.8626

Table 2: Coefficients for the Fitts' law linear model obtained with experimental data

It's important to note that although the target's distances are in millimeters this measure depends a lot on different levels of calibration of the setups. These units were used as a consistency measurement between the lamp and the web but might not have corresponded real measures during the experiments.

Finally parameter 'D' varies for every trial, meaning the distance from the shelves starting point to the target center. The final results for the application of the Fitts' law to our targets is below taking into consideration the first 16 trials where we only have one shelf allowing for more accurate results. It's important to note that although the experiment design tried to make all conditions as similar as possible on both interfaces it was not possible to assure the distance 'D' was precisely the same on the Lamp and the Web versions.

Our linear model for the coefficients seems to be a plausible one, with an increase of half a second on average from the easiest to the hardest trial and a difficulty slope suggesting that the time it takes to complete the trial grows a bit bellow the distance increment. The law uses the binary logarithm of the distance divided by the precision and in the end this results as differences of milliseconds (10ms the biggest) between our trials.



Picture 7: Fitts' law predictions for the first 16 trials

7.3.2. Cognitive (Warehouse) Task

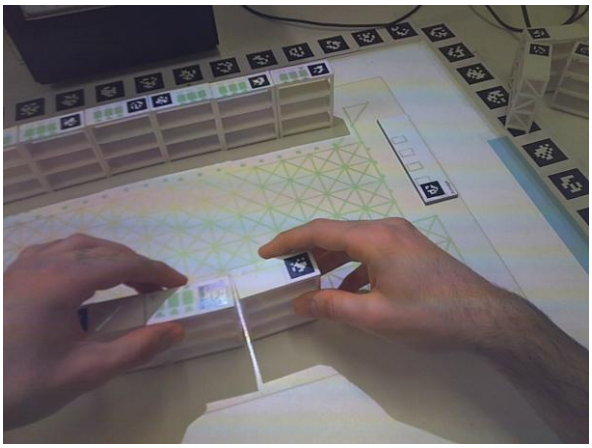
Until this point subjects shall have acquired interaction skills and also some holistic understanding of the system. Both interfaces have been assessed and the landscape of virtues and limitations is chartered in terms of human-computer interaction. One can now proceed to analyze a task that comprises meaningful problems to be approached and solved by recurring to these technologies. The idea is to now confront them with a real problem from the logistics domain and obtain a score from their performance.

The subjects were given a clear set of instructions about specifics of the warehouse functioning. They're told how to group shelves correctly, how to read if a shelf or a dock is accessible or not and to avoid blocking specific areas of the warehouse or the access from one dock to the other with their chosen shelf disposition. They are provided with 16 shelves and 5 minutes to interact with the system and try to optimize a warehouse with a specific topology according to a chosen criterion. The topology used is a simple rectangular room, 20m width, 18m height, with no walls and no inaccessible zones.

Each shelf is evenly divided into three parts, each capable of storing 3 pallets. Partial inaccessibility to a shelf drops the number of usable pallets. A maximum of 144 pallets is possible and necessary for a good score. The score computation is the average distance from the shelves to the in dock and to the out dock as this is an easily understandable concept but not that intuitive on practice. The lowest the score, the better it is.

This kind of task combines cognitive load and interaction skills and it's important to assess the not only the scores but all warehouse evolution. This is accomplished by logging every shelf movement detected along the process.

The score metric might be useful but has a serious limitation. If the subjects are unable to use or to make accessible all shelves a penalty must be applied. The problem is that there is no robust way to combine the distance score with a penalty coming from the number of pallets the warehouse is lacking. The results will be presented with the distances, the penalties and a combination of both that tries to be as meaningful as possible. For each shelf, or a fraction of it, missing the distance is increased by the double of what is missing. There is no way to calculate an average increase in distance for missing shelves because this depends on the topology as a whole. Even so this metric assures that having a few inaccessible pallets has a small impact (9 pallets, 1 shelf, 2 meters penalty) but more than this sky rockets the score to absurd values in terms of distance and this kind of makes sense because taking most advantage of the warehouse space is a first priority. Although not very scientific, some empirical tests performed by adding shelves to the worst possible positions in different layouts suggest similar increases in average distance if missing somewhat below 2 shelves.



Picture 8: Warehouse building with the TinkerLamp

7.3.3. Experts Evaluation

It would be a very limited approach to just look into this data not judging other factors of performance but to do it one must endeavor in more laborious (manual) techniques. Automatic logging makes it very hard and unreliable to gather more specific information than the scores, with the movement of a lot of shelves and users blocking the way of the camera detecting the shelves tags. To achieve a more complete cognitive task analysis I requested the precious help of two colleague graduate students, with an expertise in HCI. The idea was to have them watch the eye tracking videos in fast forward (1.5x speed which takes them a little more than half an hour for all subjects) and to rate two different metrics.

These experiments are designed to provide the subject a similar experience on both interfaces so one could expect consistent performances not only in terms of score about the main objective provided but also on common factors I called strategy and optimization. While strategy is all about building a

solution taking into consideration the premises of the score and all special limitations of the warehouse, optimization is about subjects realizing that symmetry and specific orientations of groups of shelves have a deep impact on the final output and some minor twicking considering these properties can result in great improvements.

The results from this evaluation are not supposed to be rigorously measured but instead evaluated in a 1-5 scale by both of the experts and then validated using the Cronbach alpha test. This should be a good way to extend both the insights on the cognitive task and how it relates with HCI and the overall experience of the subjects.

7.3.3.1. Strategy

The first heuristic I asked the experts to consider was strategy. Because of the limited space and the concrete requested objective our subjects should try to develop a strategy on building their warehouse. This strategy should take into consideration the size of each shelf against the total size, the average distance measurement, the shelf alignments, and so on. By watching the eye tracking videos it was clear that some subjects were really concerned about these considerations while others were more on a trial and error approach. In other words, at a first glance strategy did really seem to make significant difference, even for such simple tasks, and measuring it might help to conclude about cognitive results better than the automatically computed score.

7.3.3.2. Optimization

During the process of building the warehouse it becomes clear that, being the score derived from the average distance to the docks, symmetry plays a decisive role on achieving a good result. Also, because we're trying to fit a lot of shelves in not such a big area, soon the alignment of the shelves and the emergent pathways from their placement influence a lot the thinking and building process of the subjects. Thus, optimization can be traced by the capability of significantly improving the score just by taking into consideration these factors.

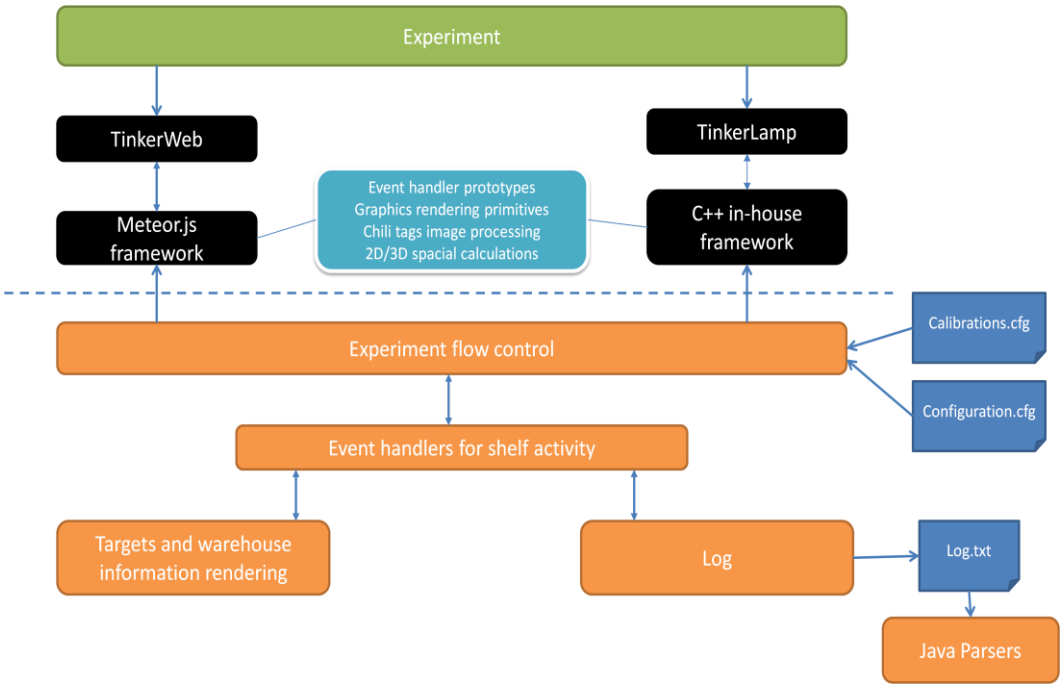
8. Implementation

The implementation of the tasks contemplated the targets rendering, detection and schemas and also the logging functions configuration. This had to be developed both for the TinkerLamp and the TinkerWeb. The Lamp software was coded in C++ using in-house libraries for events handling, like chili tag updates, rendering the targets and the warehouse information and 2D/3D spatial calculations. The TinkerWeb environment was based on metor.js, a javascript framework that has very interesting features for this project like latency compensation (crucial because of the webcam image processing functions) and data synchronization that allows the browser to handle events triggered by the tags being captured while updating the rendered abstract representations and logging all relevant data. On top of this all scripts were developed in coffee script. This is a very compact yet human readable

language that compiles to javascript and enables the reutilization of code developed in experiments, such as this one, without much documentation overhead.

The interaction tasks were developed using a logical layer over the provided framework that is responsible for rendering experience specific graphics, like the targets, and does the flow control for the whole experiment. The configuration of the flow is done by inputing a text file with all target configurations and sequencing, consisting of trial numbers and coordinates according to an internal mapping system.

Event listeners were developed to all shelves actions and call log actions, graphic rendering and flow control mechanisms. Adding, removing and moving shelves on camera sight were the basic triggers of the system. With this information associated with a timestamp, both the system's states and the procedure can be fully described and traced.



Picture 9: Implementation architecture

The 2D/3D calculations had to be calibrated to output equivalent results on both setups. The necessary parameters were obtained with empirical tests and provided through a calibration .cfg file.

The logging functions were implemented by using available functions to listen to shelf movements. The events related with the targets, like creation and hits, were also coded with specific listeners so that we fully capture the interaction process. Again, the log output has all the information needed to reproduce every step of the experimental procedure.

Picture 9 shows an overview of the implementation architecture. The orange boxes represent the modules developed for the sole purpose of the experiment. All others represent different artifacts that

supported the development. Because the architecture and the code are property of Simpliquity the details of each implementation can't be specified.

Besides the architecture, the eye gaze tracking was subject of a few tests during development. The mobility of the equipment has it's downsides as the information captured can't provide a stable, continued tracking of all tags. Fortunately, the time elapsed between loss and recovery of a tracked tag is small enough, so that crossing the data with the system's log provides accurate results. The rates were measured and optimized by reducing the number of processes running during the experiment and also by fine-tuning browser configurations. The SMI tools to render the gaze point, action zones and other features was also tested for the experiment so that one can access easily information regarding what tangibles or parts of the interface the subjects are focusing. This feature was particularly important to enable automatic extraction of data to support the experimental part on split attention effect.

9. Procedure

An important concept when actually conducting the experiment is having a script. The whole process should be thought through and performed in the same way every single time to avoid untraceable biases. We developed a script to help us guide the subjects but also to help ourselves during the experiment. It contains clear instructions for every task and all the steps we must go through. The general structure of the whole procedure is the following:

1. Welcome & general explanations about the study
2. Pre-test questionnaire
3. Calibration
4. Accustom period
5. HCI Tasks (Mod 1)
6. Analysis Task (Mod 1)
7. Warehouse task (Mod 1)
8. In between questionnaire
9. HCI Tasks (Mod 2)
10. Analysis Task (Mod 2)
11. Warehouse task (Mod 2)
12. Post-test questionnaire
13. Debriefing of participants

While most of the steps have been already explained or are self-explanatory some care for additional notes. The accustom period will be a 2 minute free-time to hang around with the interface so we avoid this kind of behavior during the first couple of tasks. The repetition of all the tasks is because of the between subjects design introduced before. The decision on having 3 questionnaires, which one shall take no more than a couple of minutes to fill, is because we want to perceive different conditions of the

subjects before and after using each interface. This allows us to account not only for their final opinions but also for their expectations, which might influence the way they deal with the system.

10. Population

Every subject is a freshman from either EPFL or University of Lausanne, 6 males and 4 females, between 18 and 21 years old, all right hand sided. None of the subjects had ever used such interfaces nor they had any prior knowledge about warehouse organization; all rated themselves above 3 out of 5 in feeling comfortable using interaction technologies. The Gender and the level of comfort with the technology the factors considered to try to make a balancing on the subjects per condition (lamp or web interface).

To name our users one might find different terminology that was just relevant for the sake of easier grouping data on early stages. While sometimes are indexed from 1 to 10, some other times a nomenclature of (S11,S12,S13,S14,S21,S22,S23,S24,S25,S31) is used. This helped the parsers to distinguish users and in some analysis remains this way with no specific purpose.

11. Statistical context

Making conclusions about a population from a sample is not a simple task and to be able to answer the research questions posed before one needs the help of inferential statistics. The next section focuses on this but it's important to provide general guidelines. A simple way to put the research questions of this study is to think of testing hypotheses. These hypotheses have to come out of some prediction and since this is a test for two different interfaces under the exact same conditions (as far as possible) one can think of all our questions from a perspective of whether the lamp and the web interfaces perform the same. This way it's easier to use and understand the statistical tests in their usual format.

Most of the tests performed will be parametric tests, t-tests, where one needs to confirm assumptions on variance homogeneity and normal distribution of the population. The problem is that the results of these verifications are not accurate for small samples like the one we have. Anyway, the parametric tests are more powerful than assumption-free, non-parametric ones so they'll be used and presented along with effect size and power of the test. Statistical power will often be under convincing levels (80%), again mainly because of such a small data sample. This represents the chance of actually finding an effect if one exists. To boost the confidence in the results provided one would have to necessarily repeat the procedure with more subjects. Constraints of different natures didn't make this possible but by providing clear results on all relevant statistical tests metrics I make sure the conclusions I draw from this small number of experiments is considered through a truthful prism.

Following the recommendations by the American Psychology Association format for presenting experiment results (Association, 2001) the t-tests will be reported alongside with effect size, statistical power and an interpretation of the null hypothesis. Because several t-tests are performed for some metrics a better visualization is possible using a table rather than the usual text reporting format

suggested by APA and I will present results this way. Linear regressions are presented in tables with the slope (beta), t-test and significance level and for variance the R^2, F statistic and p-value. All values are rounded to 2 decimal places as also recommended.

12. Experimental Results

12.1. HCI - Speed

The first simple analysis conducted concerns the interaction speed. Experience data allows for the extraction of multiple metrics that help us draw conclusions on this. One may look into the time each shelf takes to reach a target, how long it takes to first complete a trial or how long it takes to completely finish the trial. The difference between the last two is that in order to completely finish a trial the shelves must be stable on target location and not just “pass by” it. While first reaching a target is really a pure speed measurement, considering the whole trial also involves accuracy (even if we’re just talking about one shelf only it might require fine tuning before it rests in a stable position from a detection point of view).

Thus this first presented analysis, on speed, takes into consideration how long a shelf takes to reach a target. We first consider only the subjects that had no prior contact with the other interface (between subjects, n=5) and then all subjects (within subjects, n=10). We’re looking to the mean values by interface (web or lamp) for the first 16 trials (trial type 1) which, as mentioned before, have only one target to hit.

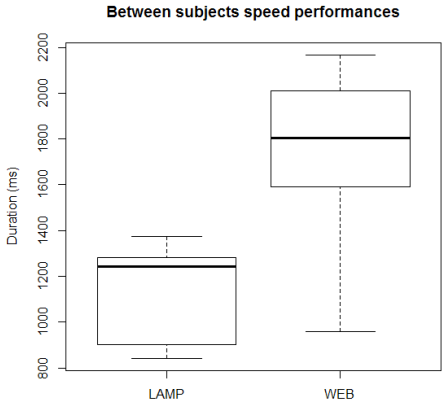
The statistical analysis validation of the data is performed using independent and dependent two-sample t-tests, suited for the categorical nature of the independent variable in study. The dependent test must be used to validate the within subject analysis (n=10) to take into account the prior completion of the experiment in another interface. The tests will be two-tailed (non-directional) meaning one doesn’t state weather a group is expected to perform better or worse than the other and results tell about the difference between the two. This means the typical null-hypothesis will be the two groups (different interfaces or different interface order) performing the same. This procedure requires some data assumptions as mention before: variance homogeneity, normal distribution of the population and independently sampled data. The latter is guaranteed by the experimental procedure itself while the others will be verified statistically using the Bartlett test for variance homogeneity and the Shapiro test for population normality.

Since the experiment was counterbalanced in a sense that every subject either used one first or the other, split half-half, one might expect no statistic relevance thus being able to conduct an independent test on the whole population (10 subjects). After conduction a paired t-test for both web-lamp, $t(54) = 5.43$, $p < .001$, (p-value = 0.6528) and lamp-web (p-value = 0.5679) one can state that there is no significant change in performance whether or not a subject has experienced the other interface before. Thus a within subjects analysis seems passable too.

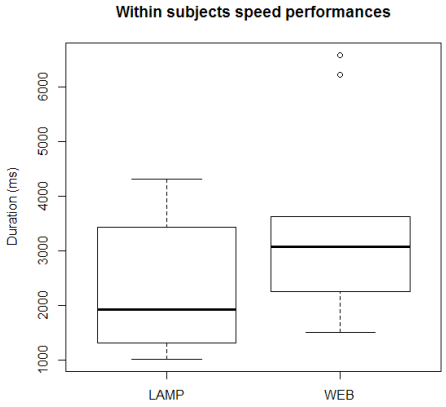
The following table presents results from all statistical tests performed.

	Between subjects (n=5)	Within subjects (n=10)
Variance homogeneity (Bartlett test) p-value	0.22	0.35
Shapiro-Wilk test p-value (population normality) - Lamp	0.23	0.24
Shapiro-Wilk test p-value (population normality) - Web	0.54	0.51
t-test t-value	-2.44	1.68
t-test p-value	0.05	0.11
t-test 95% confidence interval	[-1156.85, 2.55]	[-291.23, 2549.22]
Effect size	0.70	0.38
Power	0.13	0.19
Mean value lamp	1128.83	2296.69
Mean value web	1705.98	3425.69

Table 3: Speed statistical results



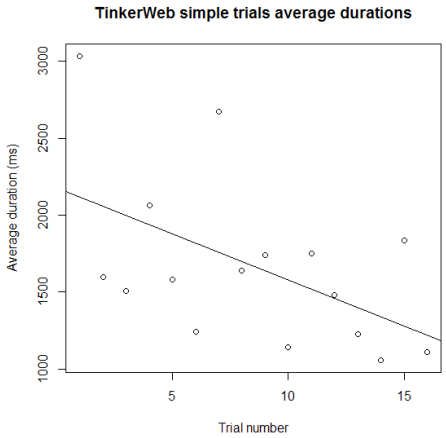
Picture 10: Speed performances of users with no previous contact with the other setup



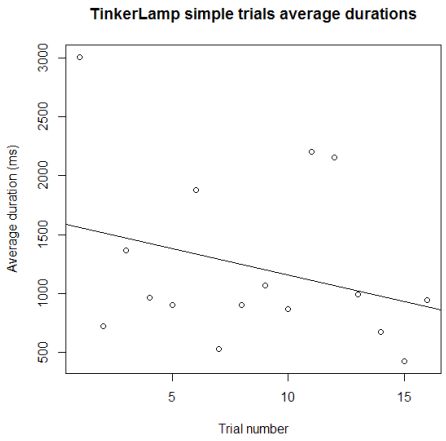
Picture 11: Speed performances of all users (both with and without previous contact with the other setup)

For the Bartlett test the null hypothesis is that the variances are equal thus one can go for a t-test with a p-value above 0.05. For the Shapiro test the null hypothesis means the distribution being normal and so with a p-value bigger than 0.05 one cannot rule it out and assume normality of the distribution. According to the obtained values all assumptions are fulfilled. The effect size is important between subjects (> 50%) but none of these tests is powerful enough (< 70%), so one is not assured at all to detect and effect if one exists. The bigger sample size could increase the power of this analysis but because we have a much smaller effect size the difference is minimal.

Another necessary verification regards the learning curves we're trying to infer from the sequential trials performed. The next plots provide linear regressions on the first 16 trials for speed.



Picture 12: Linear regression on the average speed for the first 16 trials - TinkerWeb



Picture 13: Linear regression on the average speed for the first 16 trials - TinkerLamp

For the TinkerWeb, the trial succession significantly predicted the average duration of a trial, β= -60.06, t(14)=2.30, p=<0.05. It also explained a significant proportion of the variance in the durations. R^2=0.22, F(1,14)=5.28, p<0.05

For the TinkerLamp, the trial succession didn't significantly predict the average duration of a trial, β= -45.00, t(14)=-1.18, p>0.05. It also couldn't explain a significant proportion of the variance in the durations. R^2=0.02, F(1,14)=1.38, p>0.05

Both linear regressions don't perform very well on predicting our data, especially the one regarding the TinkerLamp, show no significance. We find some significance for the TinkerWeb.

12.2. HCI - Accuracy

Now that we have conducted an analysis about the speed of our subjects we'll do the same for accuracy. While by speed one means the time it takes to reach a target, accuracy may not be as simple to define.

The first idea to capture accuracy in this experience was to use the spatial calibration properties. While a user has an interval for 'x' and 'y' coordinates in which the system considers it a valid placing of a shelf we can measure such a difference and compute a metric of accuracy from it. The problem is that interesting precision thresholds conflict with the precision of the calibration process, ie precision thresholds are typically of an inferior magnitude that calibration deviations. Because we make a calibration for each experiment and users get used to it during the interaction process one can't consider it a reliable metric.

The decision was to use time references, as with speed measuring, considering either the time it takes for a shelf to hit a target for the last time or the time it takes to complete a trial. The problem considering trials and not single hits is that we have a combination of speed and accuracy interacting with each other. Thus the approach here will be basically the same as with speed but the metric is about the time elapsed between the first time a shelf achieves a target's position and the last time it does so, ie until it rests in a stable accepted position while standing on a target.

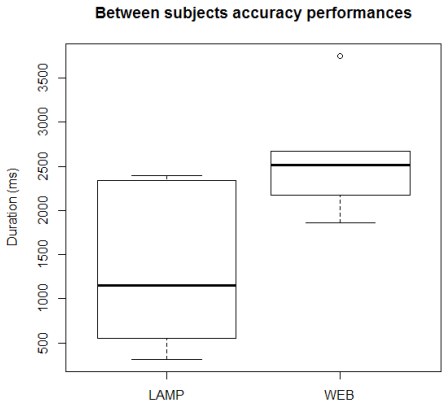
One can now proceed to compare accuracy between the two interfaces conducting the same analysis procedure as before. The next table presents data in all similar to *Table 3* but this time for accuracy. The within subjects design is passable once again, with a lamp-web p-value on the dependent t-test of 0.81 and 0.40 for the web-lamp (both way above 5%).

	Between subjects (n=5)	**Within subjects (n=10)**
Bartlett test p-value (Variance homogeneity)	0.56	0.66
Shapiro-Wilk test p-value (population normality) - Lamp	0.24	0.34
Shapiro-Wilk test p-value (population normality) - Web	0.51	0.48

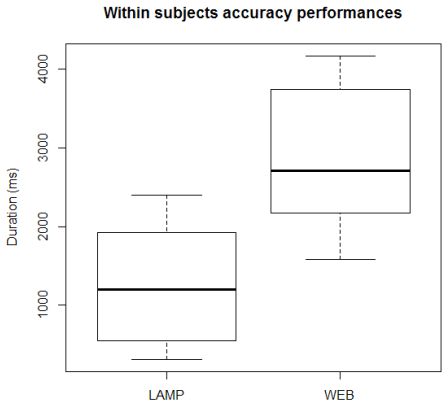
t-test t-value	-2.30	-4.25
t-test p-value	0.05	0.00
t-test 95% confidence interval	[-2512.32, 24.70]	[-2327.82, -786.90]
Effect size	0.65	0.71
Power	0.15	0.52
Mean value lamp	1351.23	1264.37
Mean value web	2595.04	2821.73

Table 4: Accuracy statistical results

Once again all assumptions are verified but significance is below 0.05 for the within subjects design. The 95% confidence interval for the between subjects design includes zero, the chances of these groups being having the same mean, even though it's by a small margin. The good news is that the within subject design even has the effect size and power necessary to draw statistically relevant and conclusive results.

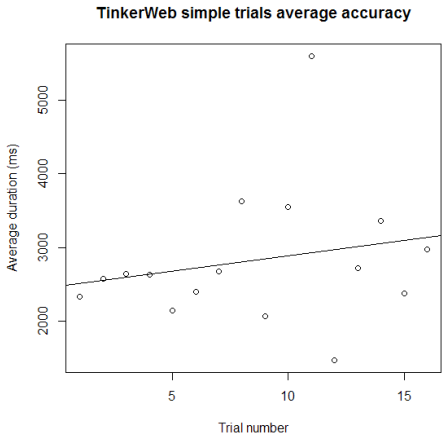


Picture 14: Accuracy performances of users with no previous contact with the other setup

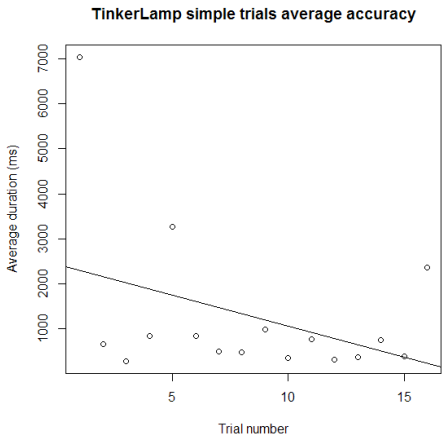


Picture 15: performances of all users (both with and without previous contact with the other setup)

Again repeating the procedure from speed on shall conduct the necessary statistical analysis on the simple trials, using a linear regression, to conclude about improvements along the attempts.



Picture 16: Linear regression on the average accuracy for the first 16 trials – TinkerWeb



Picture 17: Linear regression on the average accuracy for the first 16 trials – TinkerLamp

For the TinkerWeb, the trial succession didn't significantly predict the average duration of a trial, β= 50.66, t(14)=0.82, p=>0.05. It also couldn't explain at all the proportion of the variance in the durations. R^2=0, F(1,14)=5.28, p>0.05

For the TinkerLamp, the trial succession didn't significantly predict the average duration of a trial, β= -137.89, t(14)=-1.53, p>0.05. It also couldn't explain a significant proportion of the variance in the durations. R^2=0.08, F(1,14)=2.34, p>0.05

12.3. HCI – Trial Duration

At this point one knows what to expect about speed and accuracy but under extremely simple conditions (trials). So, shall HCI metrics just increase linearly as we scale task's complexion?

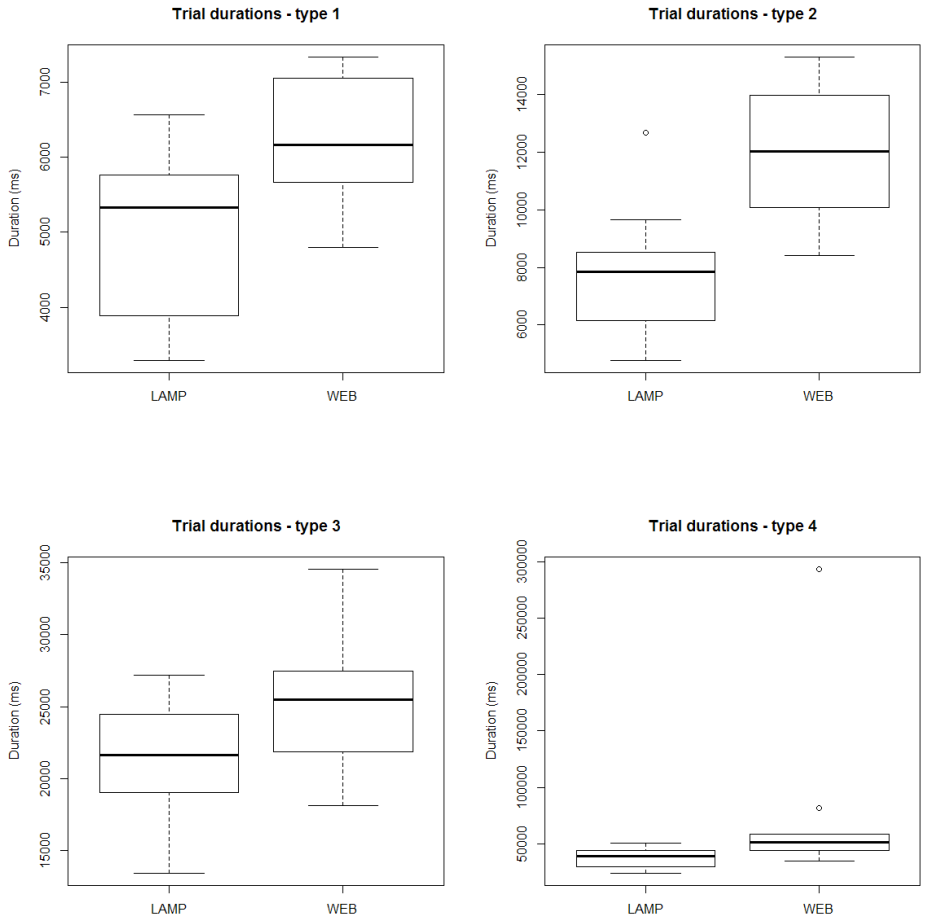
Speed and accuracy can no longer be accurately measured in more complex trial types than the ones explored before. The subjects can move more than one shelf at a time or perform fine tunings only after roughly positioning all shelves. At this point one can only rely on trial durations to conclude about overall performance on each interface.

The first step will be to verify all needed assumptions and to compute the box plots for all the trial types. The following table presents all the necessary assumptions and main results in a more compact version than before. At this point one can compare average results as difficulty increases.

	Trial Type 1	Trial Type 2	Trial Type 3	Trial Type 4
Variance homogeneity (Bartlett test) p-value	0.56	0.80	0.59	5.41e-07
Shapiro-Wilk test p-value (population normality) - Lamp	0.34	0.61	0.85	0.54
Shapiro-Wilk test p-value (population normality) - Web	0.39	0.48	0.75	6.31e-06
t-test t-value	2.54	3.94	1.98	
t-test p-value	0.02	0.00	0.06	
Kruskal p-value				0.44
Effect size	0.52	0.68	0.43	
Power	0.20	0.30	0.15	

Table 5: Statistical results per trial types

The statistical test performed give significance and effect size to the first two trials; the third one is a little beyond the accepted limits and trial type number 4 doesn't allow to state the two groups means are different. This trial type data has no variance homogeneity nor population normality and using a non-parametric test like Kruskal-Wallis the p-value is way above the usual 0.05 threshold. The small sample size and the big outliers ruin the chances for a statistically robust review of this data and the whole trial sequencing approach to HCI performance but having no power to infer a population out of this data might still leave some room to make some sense out of our subject's interactions.

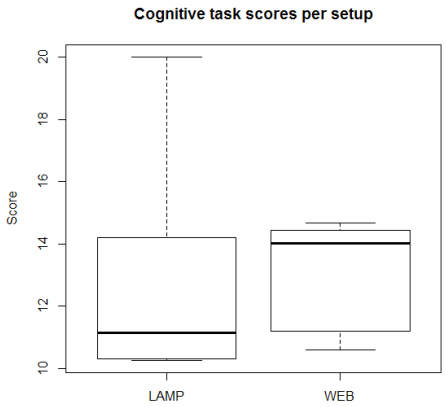


Picture 18: Trial durations per trial type and setup

12.4. Cognitive Task

The different metrics for the cognitive task that have been described before will now go through statistical scrutiny.

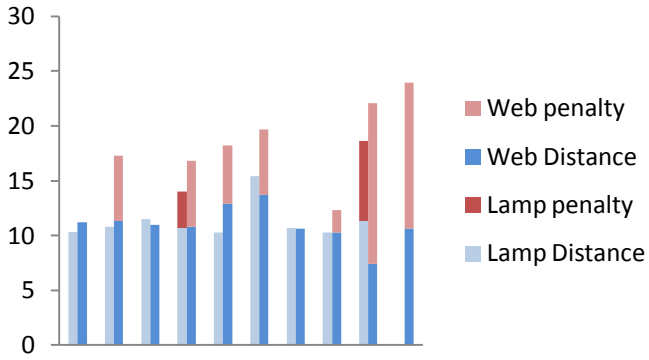
By performing a dependent t-test the way presented in the HCI section one can conclude that the means are not significantly different when considering subjects that previously had used another interface to perform the cognitive task. The next box plot presents the performances for all users on each interface.



Picture 19: Cognitive task scores on each setup

Applying the Bartlett test one can't conclude on the homogeneity of variances therefore we use the non-parametric Kruskal test. The Kruskal test with a p-value of 0.29 has no statistical significance thus these two groups could represent the same.

The chart below represents the score metric presented in the Method section, sub-divided into the distance and the penalty for not using all possible capacity. It's important to remember that there is no fully integration between the score and the penalty because of no robust model to merge the results on different scales (distance and capacity).



Picture 20: Score metric results and penalties for both setups

These results obtained represent the score metric presented in the Method section, sub-divided into the distance and the penalty for not using all possible capacity. It's important to remember that there is no full integration between the score and the penalty because of no robust model to merge the results on different scales (distance and capacity).

The best performing subjects were capable of reaching an optimal score around 10 meters with no penalties. It's important to note that only two subjects we're not able to use all shelves properly with the Lamp interface but this number rises to 7 when with the TinkerWeb. Subject number 10 has no

lamp performance because during all the performance shelves we're blocking any possible path from one dock to the other. While a small score is good, on in this case it means the worse possible outcome.

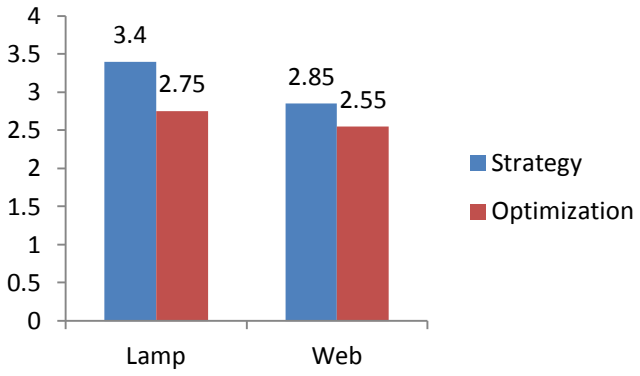
12.5. Experts Results

The requested experts for further evaluation of the cognitive task performance considered both strategy and optimizations. The first step is to validate their evaluations. To do so, by pairing up all the classifications from both experts, one can compute the Cronbach alpha:

$$\alpha = 0.87$$

This result is above the recommended 0.7 for confidence of the results through average correlation between their evaluations. Thus one can consider a quantitative analysis on strategy and optimization as a way to substitute the unsuccessful automatically collected data analysis.

The next chart presents the results by interface for both of our criteria. All scales were considered from 0 to 5.



Picture 21: Experts average evaluations per setup on strategy and optimization for the cognitive task

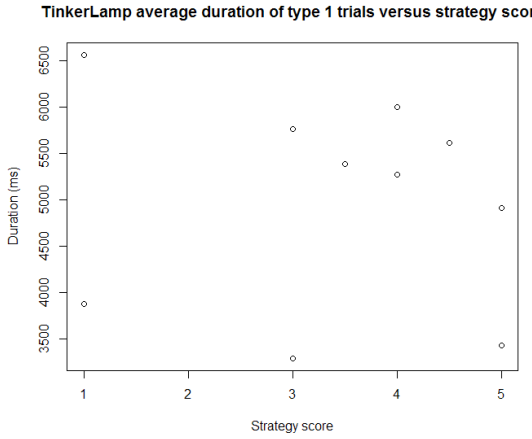
12.6. HCI VS Cognitive task results

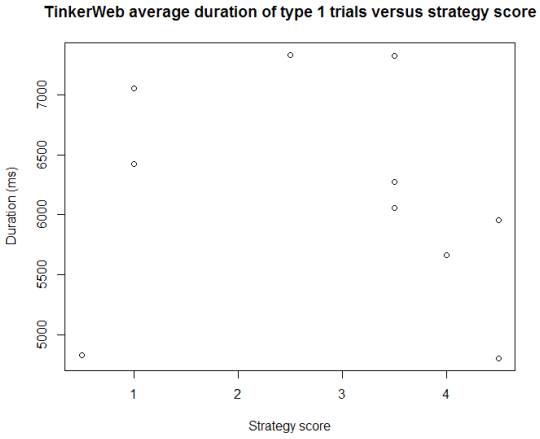
The results from both the HCI and the cognitive task can be crunched together to draw a deeper insight on the differences between the setups. Because the cognitive scores results were inconclusive this exploration will be limited to the comparisons of both HCI tasks durations on trial type 1 and all trial types against cognitive evaluation by experts only.

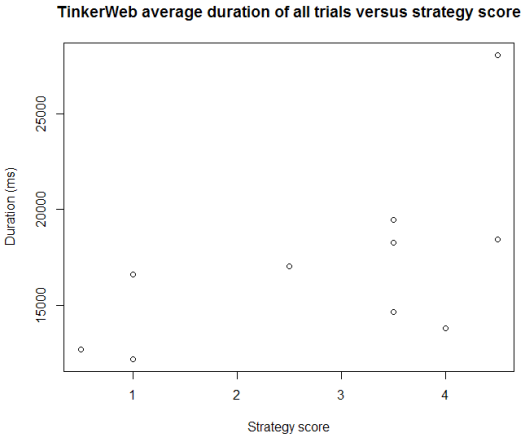
One can assess each one of our participants average durations against the strategy and optimization scores provided by the experts. Checking the correlation coefficient between the two variables will quantify the way in which the two variables are related.

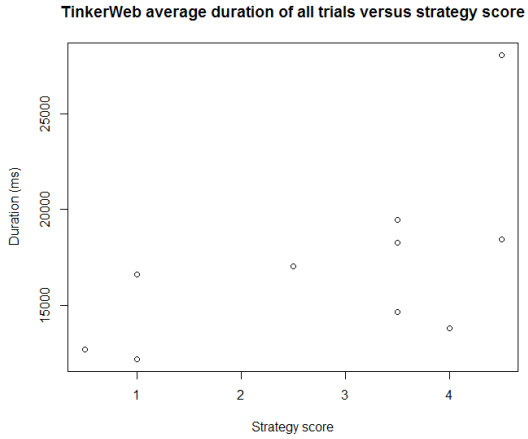
	Trial type 1	All trials
TinkerLamp Strategy	-0.13	0.31
TinkerLamp Optimization	0.12	-0.10
Tinker Web Strategy	-0.15	0.60
Tinker Web Optimization	-0.45	0.62

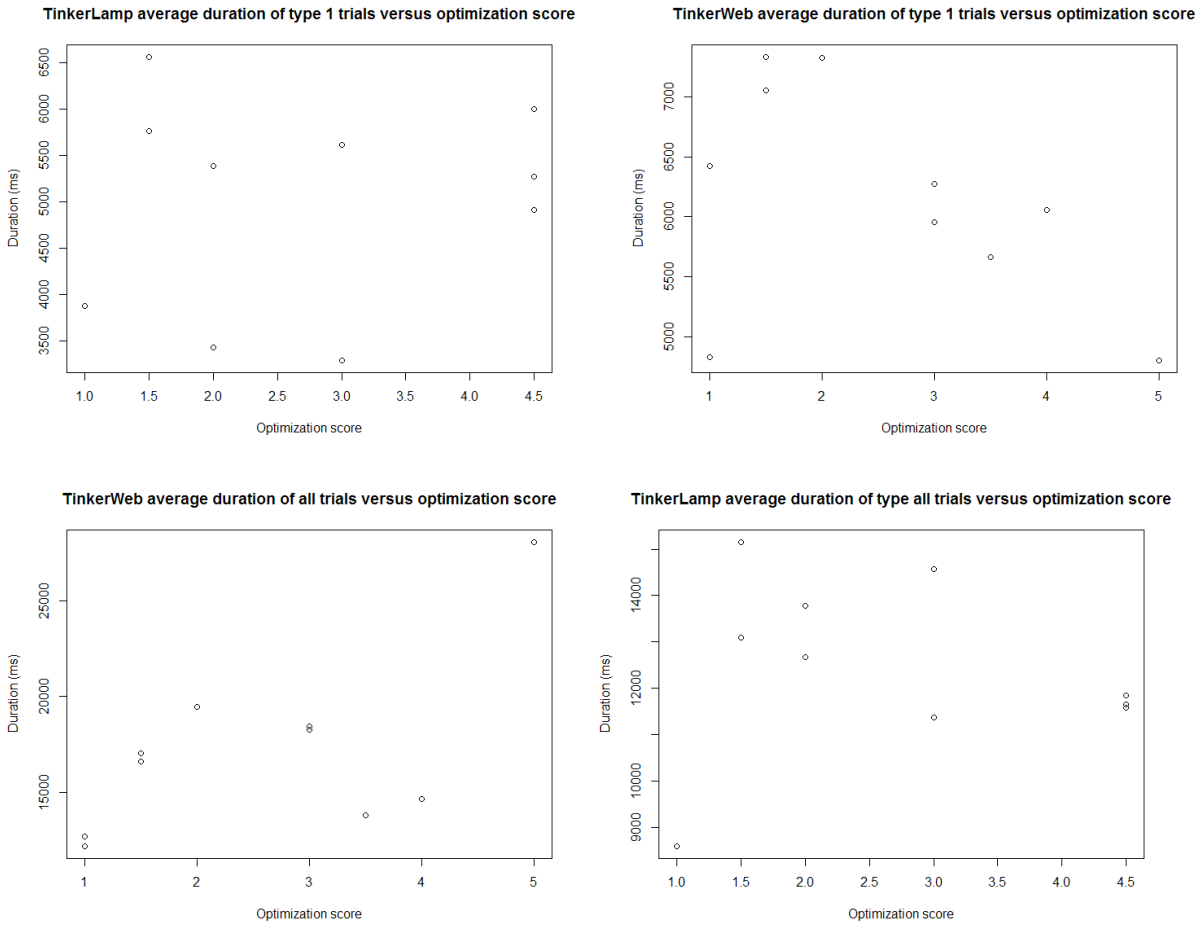
Table 6: Correlation coefficients for HCI tasks duration and experts analysis results











Picture 22: Plots for all the different correlations tested between HCI tasks and experts analysis

Correlation values are mostly insignificant and this is easily understandable by plotting the results. The only results with some relevancy concern the TinkerWeb for all trials on both strategy and optimization, with a positive correlation above 50%.

12.7. Mediatory Variables

12.7.1. Expertise grouping

Another interesting analysis would be to understand how well our subjects deal with these setups and compare this to their results. Through the questionnaires one can picture how well they expect to perform and this might have an influence in the performances. This influence could tell us about possible evolutions after continued usage or how some people might be ahead of others, an important factor in the educational frame.

Subjects were asked how comfortable they feel with these technologies using a Liker scale. Even if asking a general question about how comfortable they feel with this type of technologies, their

answers could be a good indicator of how they will perform. Giving a low value on this question means above all things that they don't feel confident, not comfortable, but that should be enough.

Out of our 10 subjects the ones giving a 4 or 5 answer (1-5 scale) were considered experts (40%). A correlation test between the self-elected experts and their cognitive performances suggest no significance at whatever level considered.

Correlation coefficient = 0.08

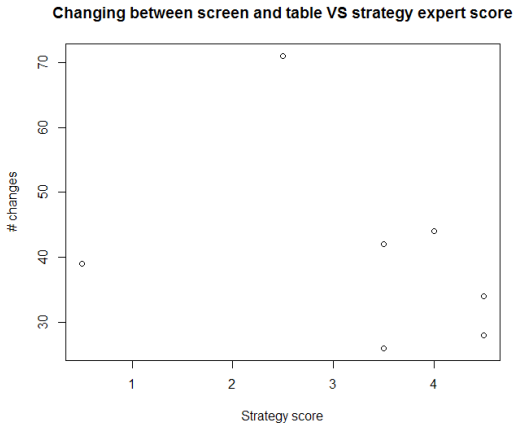
12.7.2. Split attention effect

One of the most relevant parts of this study regards the existence of a split attention effect with our new setup, the TinkerWeb. One of the biggest concerns of separating the representations between the screen and the table with the tangibles is that the users will somehow drop their performance indexes because of a high frequency of focus change (Ayres & Cierniak, 2012).

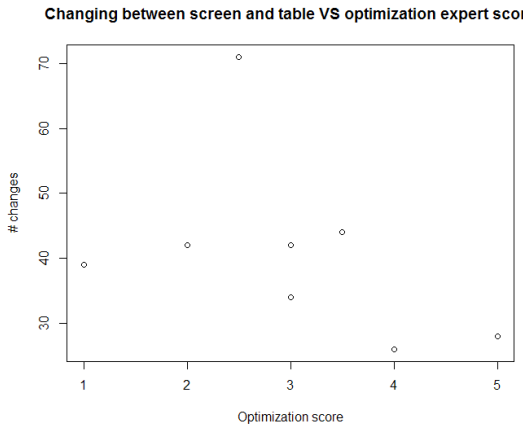
Comparing the cognitive results with metrics that can somehow reveal this effect is the purpose of this section. Because our cognitive task measurement had non-significant results one must recur to the experts' evaluations. The specific metrics considered for this part of the study were the number of changes between the screen and the table and the time spent on each one.

The next table and plots present the results for crossing this data. It's important to note that two of the subjects eye-tracking provided useless data (S22, S31). The eye-tracker is not very stable for some facial structures and sometimes these results in an abnormal tracking of the pupil. Thus these results were only considering 8 out of out 10 subjects.

One would expect to find some correlation between the cognitive performance and the cognitive results on the TinkerWeb if the split attention effect plays a significant part in performance.



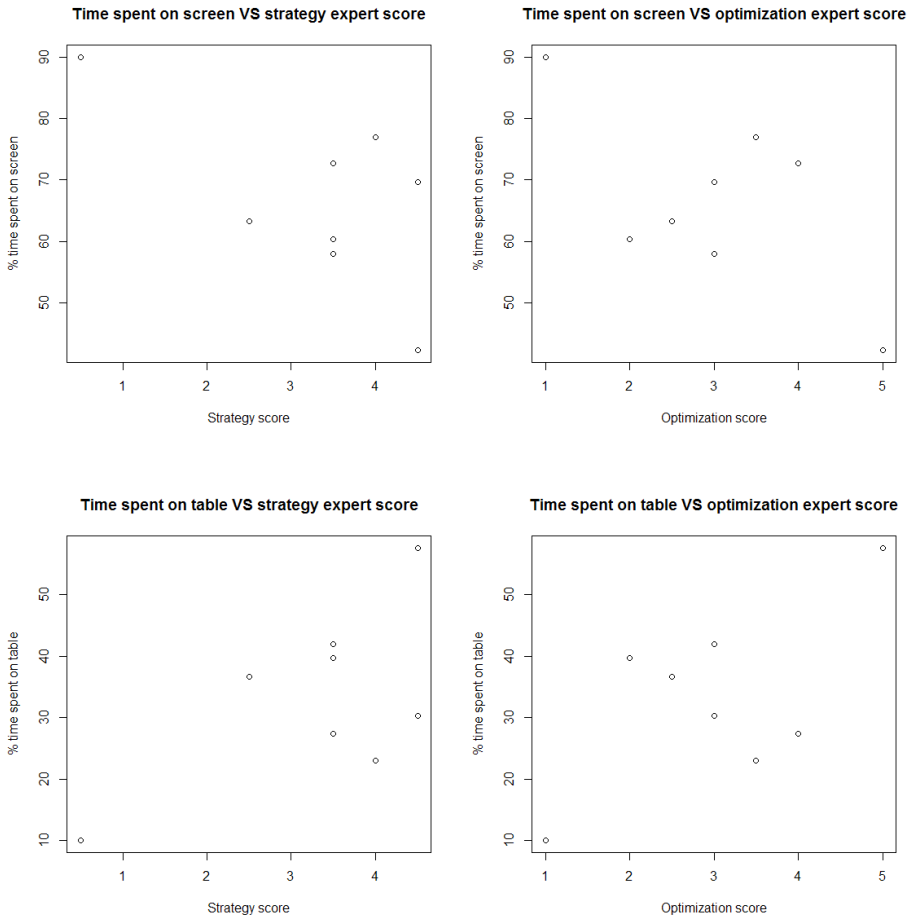
Picture 23: Strategy scores against number of changes between screen and table



Picture 24: Optimization scores against number of changes between screen and table

The correlation factor for the experts evaluation on strategy against the number of changes between the screen and the table was -0.34. For the optimization metric we get a correlation of -0.43.

Now it's time to compare the amount of time spent in each part of the interface (computer screen and the table with the tangibles) against the same cognitive performance measurements of strategy and optimization.



Picture 25: Time spent on each part of the interface VS cognitive performance metrics

The correlations for these four plots are respectively -0.64, -0.62, 0.64, 0.62. The opposed results are expected because the time spent on one interface is opposed the time not spent in the other one. There is some significance to these results as one gets negative correlations above 50% for time spent on screen and the inverse for time spent on the table.

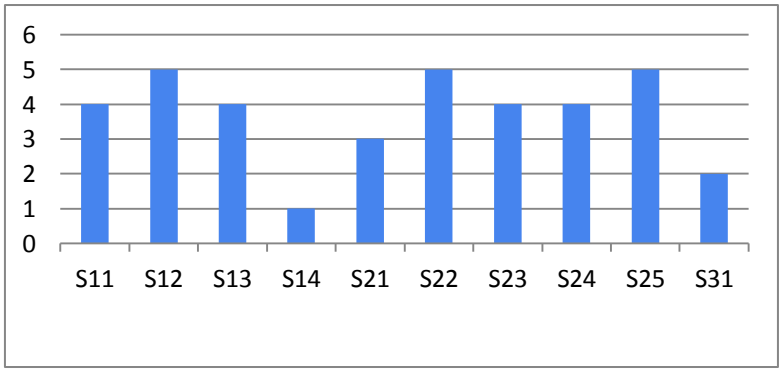
13. Discussion

The last section was about getting all possible statistical confirmations out of the collected data and at this point one has enough information to give answers to some of the research questions, state others as inconclusive, and draw interesting insights out of the different levels analysis performed before.

I'll introduce each sub-topic's discussion with the relevant results of the questionnaires conducted. By crossing the answers to the questionnaire with the conclusions being taken from the data one might get a confirmation or a misperception by the subjects that could be worth of a note. Then I'll discuss the obtained results and how strongly backed up they are by the statistical analysis.

13.1. Speed

Starting by the questionnaires, a quick scoop can give a hint on what to expect on this so that's what we'll do for starters. From the following chart one can tell that most users find the Lamp faster than the Web but being the average 3.7 (out of 5) one might not expect a dramatic difference between the two.



Picture 26: Questionnaires results on speed performance

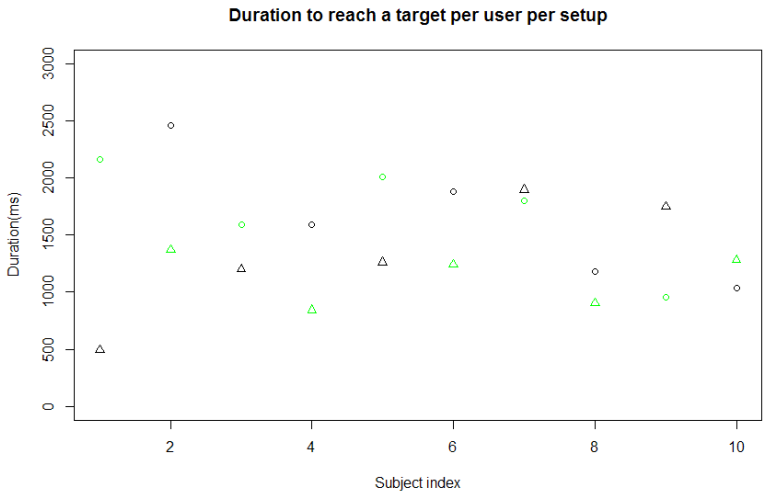
The analysis performed before showed a median about 0.8 seconds faster than the web's with statistical significance. If one looks to the mean values, the difference is about half a second. The mean and the median have different properties, being the former better to judge normal populations like this one but being less robust to outliers than the latter. Here we can see they tell us almost the same. While there is a remote chance of the web interface performing almost as good as the lamp we can expect users to have an approximate 1 second difference between the two if by looking into what the quartiles tell.

Considering all the subjects, by ignoring they have conducted the same procedure with another interface, statistical relevance is dropped (at the limit) but still it suggests a very similar average

difference, this time around 1 second. This increase in sample size should lead to clearer results but because of the within subject design all one can do is to refrain the importance of a bigger population size. The upper bound of our data through this view is represented as outliers (the dots instead of the whiskers on top) for the web interface. This is easily understandable by watching the performance of these specific subjects and see that in some trials they took way too long to adapt to moving the shelf without looking directly to their hand and doing it by the video background on the screen instead. In this case of having outliers the median might be a better reference than the mean but as one can tell from pictures 9 and 10, the difference between medians is pretty much the same as for the mean values. Picture 10 also shows the interesting fact that taking into account the users that have previously performed the trials on the other interface increased considerably the average times for both interfaces in a consistent manner. One would expect that having performed the same tasks before, even if using another interface, such would tend to produce better results but that's clearly not the case. Observing the videos one might tell that users try to mimic their interaction from one interface to the other but end up having some trouble adapting to a new reality.

It's quite difficult to assess speed purely once we scale to trials with multiple targets, as users adopt different strategies that might condition the efficacy of the metrics. Therefore I leave the analysis of the more complex trial types to a later stage, when looking at trial durations and not specific target metrics.

At this point there's a clear picture of which interface is faster and the magnitude of this difference. We'll now look at a bigger picture, at what happened to each one of our subjects. This helps to identify possible causes and effects. The following plot presents for each subject the same measurement: how fast a shelf hits a target. The circles represent the web interface and the triangles the lamp. If a circle is green it means that was the first interface tested by that user.



Picture 27: All subjects average time to reach targets, on each setup

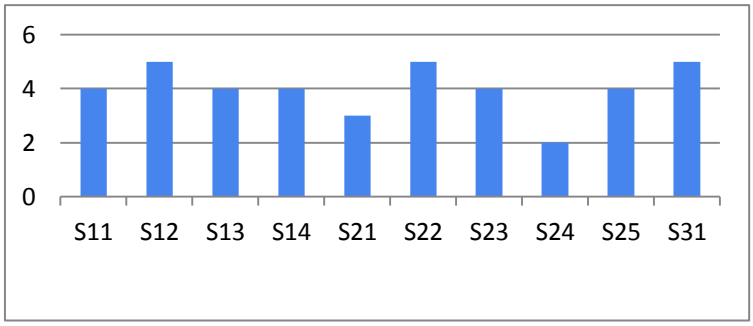
This disaggregation of data makes it clear for some of the results obtained before. There is no pattern between performances, preceded or not by experience on the other interface. Yet an interesting case, worth exploring, pops up from this different angle. How could subject seven perform better with the web interface? By carefully watching his eye tracker log one can tell that this subject had a great difficulty with the lamp because of shadows, casted by the shelves and himself, while fine tuning shelves to the targets. The problem here is that some specific positioning of the hand holding the shelf can block the projection of the target on the table making the task a lot more difficult. This subject identified his problem and corrected it but not in time to make a stronger performance than with the web interface where one finds no such problematic. This particular problem highlights the fine contours of these complex interfaces and how difficult it is to test for and survey all relevant factors.

the request for being “as fast as possible” with very different attitudes and there is not much one can do about it. Even so our best web performance is among the best lamp results and that particular user actually had some trouble with the lamp himself. Thus it seems possible to be better off with the web interface but that’s not what you’ll expect to see.

After looking at our aggregate results and dissecting those into each subject’s one must still look at the performance evolution of the subjects with repeated completion of trials. The results are quite interesting as one can clearly see from the charts 11 and 12. By calculating the average time for each trial and recurring to a simple linear regression one can see an improvement on both interfaces. The lamp interface starts off much better than the web, almost half a second below, but the improvements in the web interface are bigger and by the time of the 16th trial we can see that the gap is quite smaller (around 0.25 s). The web interface has to deal with users coordinating their movements while looking indirectly at their hand (through the screen). This is the kind of process one could expect to improve quickly after a couple of interactions and this is likely evidence of such. The oscillations in speed are understandable because we’re dealing with very short time measurements and we don’t account for reaction times and other minor factors that could impact these millisecond measurements. Only the TinkerWeb regression was significant assuring an improvement over repeated trials but one can observe a similar tendency on the TinkerLamp.

13.2. Accuracy

Repeating the same procedure used for the previous section one may start by a glimpse into what subjects felt about the interfaces. The following chart presents even more conclusive results than before: no subject found the web interface faster and with an average of 4 out of 5 one might expect slightly more pronounced results for accuracy than for speed in favor of the lamp interface.

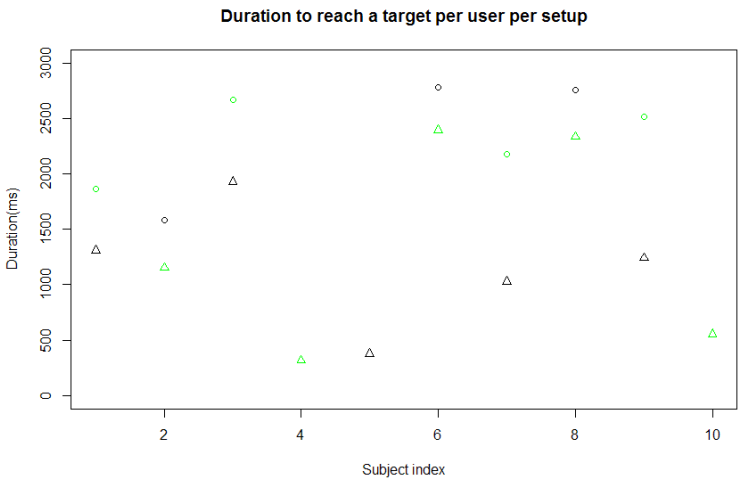


Picture 28: Questionnaires results on accuracy performance

The web worst performers as outliers were something experienced before and here the same problem appeared: some subjects really get into trouble while fine tuning in the web interface. If we look to the mean and the median they again perform quite similarly and the accuracy difference seems to be situated around 1.3 seconds. Anyway, statistically we saw this information is not very interesting.

Considering the statistical relevant analysis one can state that the lamp performs clearly better, about 1.5 seconds. This time around not even the best web interface performers could get close to the lamp numbers. While some problems may arise with the lamp shadows and hand positioning holding the shelves the web is consistently worse and a couple of insights on this arise from watching the eye tracking: computer screen latency induces users to keep moving a bit beyond the right position and because these latencies are variable it's not easy to develop a strategy to cope with it. Another reason might be that the small movements to fine tune are difficult to track in the screen and because users are not yet used to this indirect form of manipulation they make systematic mistakes of moving their hands in the wrong directions.

Again one can drill into each subject's performance to better understand what kinds of behaviors occur in specific and again no patters from previous performing the tasks are present.



Picture 29: All subjects average time to fine tune each target, on each setup

Finally the same approach to the improvement over usage can be used. With simple linear regression models over the first type trials we can see how subjects evolve their accuracy on both interfaces. With a model that has a very low significance on all cases one can still observe a slight improvement on accuracy with time for the lamp. The web interface however seems problematic because subjects seem to be less accurate along the time. If they are more experienced, how could they be getting worse? Again recurring to the videos a possibility arises: latency. Although almost imperceptible, the latency inducted by the usage of a low-end camera and an in-browser image processing seems to exist. Of course users would be able, probably subconsciously, to perceive and account for this latency but a careful analysis suggests the processing delay is not fixed and so the dynamic behavior makes it much more difficult for users to account it. At this point there's a reason for at least no improvement but not for worsening accuracy along trials. The answer here seems to lie again with motivation and with how the frustration of not being able to quickly perform a simple task ends up worsening results. Anyway, these are no more than insights one could capture from attentively watching the performances with the eye tracking captured video but no scientific procedure was applied to further confirm such theories.

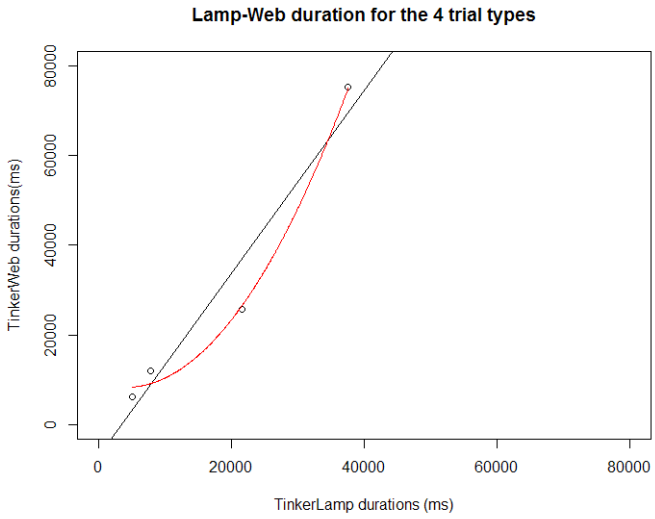
Again one can conclude an advantage for the TinkerLamp but of a small magnitude. The performance evolution also seems to favor the TinkerLamp but the low significance of all results suggests both interfaces experience a relatively small improvement over time.

13.3. Duration

Before we draw any conclusions on trial types two, three and four there is still one interesting analysis that hasn't been performed. When designing the experiment variations using translations and rotations were used on the first trial type but this wasn't taken into account until this point. Looking back on the plots provided one can easily conclude these simple changes have no apparent effects. The changes in rotation and translation forms were performed in intervals of four trials but our results present no signs of any manifestation on performance because of these modifications.

All the results seem consistent with what has been seen before, confirming the supremacy of the lamp interface, but no other results are clear from this point of view. Because one can no longer clearly separate the movement from the fine tuning, the standard deviation between users can be insightful about the error margin. Most groups present very regular deviations. The web interface for the most complex trials has some problematic outliers and by the videos one can understand how the latency of the system affects performance with more shelves being used.

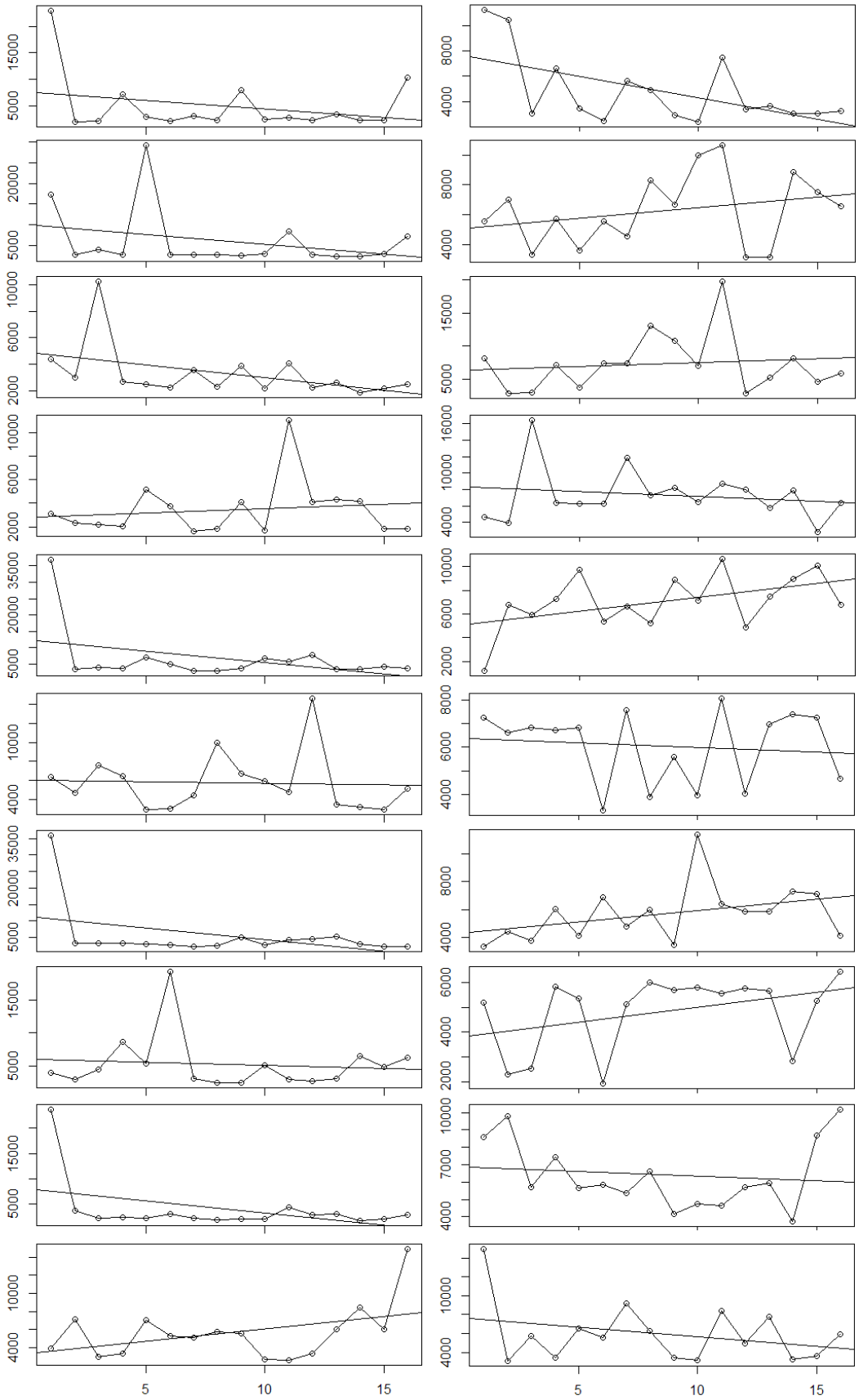
After this analysis on all trial types one can try to answer the complexion scalability question posed before by plotting web mean values on the four trial types against their homologous from the lamp.



Picture 30: Lamp and Web trial types durations models

There seems to be a slope of around 2, meaning with increasing complexity the subjects take the double of the time they need in the Lamp to perform the same HCI task on the Web. This is a very basic approach and because of the slightly smaller slope for trial type 3 and the fact that trial type 2 consists of only 5 trials while trial type 3 of 10, one could say that a quadratic model would fit better this data, representing a more typical escalation on this type of complexity situations. The previous chart presents both a linear and a quadratic model for our data and a confirmation that complexity seems to grow like a quadratic progression when taken from the Lamp to the Web interface. By the time we have 10 shelves being used, the type 4 trials, users are taking 40 seconds with the Lamp on average and 80 seconds with the Web. One can tell that as we scale on complexity, interaction becomes much more difficult with the TinkerWeb.

Before calling to an end the analysis of trial durations one can now revisit Fitts' model idea to this experiment and try to make some sense out of our data predictions. The next series of charts represents durations to complete each of the first 16 trials, per interface, for each one of the subjects. A simple linear regression line is provided for each of the plots but no exhaustive statistical analysis was conducted for all the performances.



Picture 31: Lamp on the left; Web on the right – Duration (ms) of the first trial type for all users

At a first glance it just becomes clearer what had been concluded before: almost all subjects improve regularly their performances with the Lamp but with the Web this is not clear and maybe for the reasons that have been pointed out. But these values allow for a direct comparison with Fitts' law predicted durations. The Web values fluctuate too much and we've seen their way above the predictions mainly because of the accuracy component. The lamp however presents more "well behaved" data. The values are much more stable and closer to the predictions. We've seen some reasons why accuracy might get values higher than expected but besides pointing out that accuracy on both interfaces can be considered as an impact factor some patterns seem to interline some subject's performances. Still, the huge fluctuations that we now know are mainly due to accuracy issues don't allow predictions for any interaction based on such a linear law. Also the improvements over repeated trials we explored before didn't have a clear effect on predictability. This is not good, especially for the TinkerWeb which relies on a necessary familiarity with the setup.

A concept introduced in the late 19th century by Sir Francis Galton (Galton, 1886) might have something to do with these oscillating performances. Regression to the mean is about explaining how extreme values tend to return to the mean value because of the part luck plays in repeated performances. Here one wouldn't call it exactly luck but the concept remains as we have a need for accuracy that our experiment proves hard to control and improve, and the speeds are very reliant on the engagement and attention levels of the subjects. Under this light seems clear that lamp performances can be a simple case of regression to the mean (the oscillations) with some improvement of course. Now even if the Web data seems to have more variables not accounted influencing performances we should at least expect to find a similar type of pattern. At first sight this is not obvious so running a one-side moving average one can have a clearer picture as the procedure balances performances along trials with previous results.

After performing moving averages with window sizes from 1 to 3 (more becomes irrelevant) the results are interesting: "fluctuation" is still a fact. As regression to the mean doesn't seem to play a role here and this is good evidence that the influencing factor besides skill must be quite dynamic, just as the latency issue supposed before. Unfortunately only with more specific testing with subjects one could arrive to more decisive results.

13.4. Cognitive score

Overall we concluded that there is no significance in the data collected. The analysis provides some rough ideas of what might be happening but no evidence can be drawn. From the start there was a high risk of obtaining irrelevant results from this approach and here is the confirmation. This was a very objective experiment with a clear score to make comparisons but unfortunately, on both setups, not all our subjects could achieve a performance where one can compare the scores objectively. The penalties introduced were a way to account for the sub-performing subjects without having to completely discard their results. The score was supposed to measure all shelves on context but

unfortunately some subjects couldn't get to this level. With the small population one had for this experiment this would inevitably result in results that lack significance and allow only for some insight.

The questionnaires were not considered for further sections of the discussion because although subjects have a good perception of their interaction performance it doesn't apply to the cognitive task. Their understanding on performance and their evaluation is directly tied to the score obtained and other metrics measured by the eye-tracker, making obsolete the exploration of their opinion and expectiations.

As a rough perspective one can tell that users, again, performed better with the Lamp but the differences where again quite small. The most notorious feature of this part of the experiment is the huge penalties we can see in most of the Web attempts on the cognitive task. Looking into what these users did one can tell that there was some trouble with the abstract representations on the TinkerWeb. The shapes and colors used to flag accessible or inaccessible shelves, capacity and available paths might be too complex making the whole image on the screen difficult to interpret. The augmented reality and separation of the abstract representations among the physical objects might be an advantage.

Another curious fact was that a couple of TinkerWeb users could quite easily perform as consistently as the best Lamp attempts on building the warehouse. One of the users provided a very insightful answer to this on his questionnaire by telling that he would use the screen only as kind of "minimap" of the construction while focusing on the physical objects to arrange and manipulate the warehouse. This way he could take advantage of the tangibility for realizing proportion, symmetry and other important factors while having on the screen all the meta-information necessary to guide his process.

13.5. Cognitive experts

Subject performed better with the Lamp on both metrics. However the expert evaluation suggests a bigger gap while developing strategies than while performing optimization between the studied interfaces. Expert evaluation detected a 19% difference concerning strategy and an 11% difference in optimization. These results are quite unexpressive and a solid point in favor of the TinkerWeb. If performance improvement over repeated trials was a reality for this type of task as it was with simple interaction the TinkerWeb might perform at the level of the TinkerLamp after some familiarization with this more complex process.

One factor that might explain this bigger issue with strategy might come from something identified before. The difficulties with the abstract representation on the TinkerWeb complicate the cognitive process while building the warehouse. If the users don't have a clear picture of how they are performing, which shelves are available and what path's are blocked or not it's much more difficult to develop a strategy. Instead, as one can tell by watching the videos, subjects seem quite lost at some points while building the warehouse with the TinkerLamp and only realize some of the constraints they

are facing after placing all shelves and realizing they don't have enough space or that there is path between the two docks.

13.6. HCI vs Cognitive

The comparison of the interaction task values with the cognitive task values only gives some significant data regarding the TinkerWeb. If we look at all trials in the interaction part, a good performance is associated with a good score in the cognitive task. This makes sense because the more complex the trials get, the more they somehow relate with what is expected while building an entire warehouse.

The Lamp non significant results should be associated with the fact that during interaction the basic manipulation becomes automatic for the users. During the cognitive task they can fully concentrate on the complexity of the task and not in interaction. Thus the good and bad performances are probably not at all related with interaction issues, contrasting with what happens with the TinkerWeb. Nevertheless these are results that our experiment was not designed to measure and the only concrete result of the comparison between interaction and cognition is that there is an effect relating the two in the TinkerWeb, suggesting that it's really important to improve on the interaction issues if one wants better results at a cognitive level.

13.7. Expertise and Split attention effect

The expertise approach gave very poor results with a very low correlation coefficient. It seems that no matter how comfortable, or confident, our subjects feel about the tasks they are about to perform, results won't match such expectations in any systematic way.

Concerning the split attention effect, and before going into the cognitive results against the eye-tracker measurements, it's important to explain why no such study was conducted with the basic interaction part of the experiment. The results would probably meaningful now that we know how much fine-tuning (accuracy) affects performance, especially on the TinkerWeb. The thing is that this is an intrinsic part of the setup and although there is an obvious split attention effect the impact of it should be only a matter of interaction. The cognitive task is the one where one is interested in comparing the split attention effect and the results are even more interesting because one saw there is no clear impact of the interaction aspects on the cognitive performance.

Now focusing on the cognitive results provided by experts and the split attention metrics provided by the eye-tracker, the results showed that performance on strategy seems to improve with a reduced number of changes between the screen and the table but the significance is way too low to conclude anything. On the optimization part, the lower scores seem to tell nothing but one can see a tendency on achieving better scores with a lower number of changes. Evidence has again little significance due

to the sample size. Both results have low significance but point the same way: a better performance with fewer changes confirming some sort of split attention effect.

The most interesting results concern the time spent on each part of the interface and not exactly the number of changes from one to the other. With correlations of about 60%, one can state with some confidence that cognitive performance is better the more time a subject spends looking at the screen. This applies both for strategy and optimization, with very close results, suggesting there is split attention effect that affects both dimensions.

These results, confirming somehow the existence of a split attention effect, and confirming the screen representation as the most important for the cognitive process raise some concerns for this setup. We know that many advantages of TUIs, explained in detail earlier in this document, are comprised to the table representation a it's positive effects might be endangered by this need to stay in contact with the abstract representation on screen.

14. Conclusion and Future Work

The unquestionable advantages on using TUIs on learning contexts and the recent breakthroughs on supporting technologies capabilities can help to redefine some of the paradigms of education systems. The identified issues are real problems and the solutions designed so far have addressed them quite well but, as we try to reach a broader public, some features must be dropped and thus some previous results must be questioned.

This study will help us to understand how the different dimensions on the proposed technological setups can influence the interaction dynamics, essential to the learning process itself. Being able to confirm our hypothesis with the data we collect might not only tell us how suitable these new solutions are but also what changes are going in the right way and what needs to be considered or reconsidered on future iterations of supporting technologies evolution.

The questionnaires have shown that the users get a very accurate feedback from both interfaces because the answers to the questions regarding both interaction and cognition were a very good reflection of the experimental experience. This is important not only to better validate our data but also as evidence that we can expect meaningful feedback from their overall experience.

Concerning speed, the TinkerLamp performed slightly better than the TinkerWeb with the latter obtaining stronger results on performance improvement.

Concerning accuracy the TinkerLamp performed better than the TinkerWeb but both setups are difficult to improve upon on this dimension. This should be a concern in the future for it's important that users can overcome these interaction issues so they can be concerned with higher abstraction levels of the system.

Trial duration confirm what we concluded from both the speed and accuracy: a consistent better performance from the Lamp. But here the results have a deeper impact. The quadratic relationship between the two setup durations per trial suggests that, as complexity increases, the Lamp over performs the Web quadratically.

Our prediction values obtained with Fitts' law suggested very little performance differences for the different trials, far from what was obtained. Accuracy was a very unpredictable factor. The TinkerLamp got more stable results when compared to expectations but still one can tell that this type of prediction that works well with Graphical User Interfaces (GUIs) is not suited for TUIs. It has been argued that linear models don't perform well on long distance motions like most interactions with TUIs (Langolf, Chaffin, & Foulke, 1976). Maybe in the future a new kind of predictive model can be adjusted to this kind of interaction or maybe the issues suggested by this work, namely dynamic factors such as latency, can be improved making a linear model viable to predict interaction performances.

Cognitive task was not successful in significantly measuring differences but allowed to identify, again, a small advantage to the TinkerLamp and also an issue with the complexity of the abstract representations on the TinkerWeb.

Relating interaction with cognition presented some significant results in the TinkerLamp, suggesting that speed and accuracy, especially as the interaction gets more complex, influence the capability of dealing with the cognitive endeavors.

The predicted experise of our subjects told nothing about how they perform. The split attention effect couldn't be significantly confirmed, although there was some evidence of it, opening the doors to something meaningful. The time spent on each part that tells us something with significance: spending more time in the abstract representation provided by the computer screen generally means better cognitive performance. Both parts of the interface are complimentary but if too much time is spent looking at the screen all the positive impacts of the tangibility of the interface might be compromised.

Generally the results point out the performance differences between the TinkerLamp and the TinkerWeb but unfortunately the significance of most of them were below what's desirable in an experiment of this kind. A bigger population would have been crucial to draw more definitive results but the resources for this weren't available. The change of strategy by Simpliquity during the experiments restrained the investment in a more powerful study but the results collected were still passable of an interesting scrutiny.

In the end, one has a couple of clear characterizations of both interfaces and some good ideas on where to improve and what to be investigated on further experiments. Different dimensions, of importance in the sphere of TUIs, such as exploration, fun and collaboration, could be approached for both the setups providing a broader comparison. Still one could say the most important aspect to be further investigated would be somehow related to the dynamic factors that seem to play a big role on the whole experience with the TinkerWeb. We know that working with lower-end technology and browser processing is resulting in erratic latency and other kinds of phenomena that interfere with a

more natural usage of the equipment. Also, the split attention effect seems to play a decisive role on the cognitive performance but spending most of the time looking at the computer screen neglects the important features of the tangible interface and defeats the purpose of the interface.

From a business point of view one can state that this product has potential but it's not ready to be a classroom essential. The confirmation of this fact is that, as this study comes to an end, Simpliquity is changing their strategy dramatically. Students need reliable, flexible and straightforward technologies that help them. A stiff setup might require too much focus on itself rather than the concepts and abstractions that it should seamlessly convey. The improvements suggested before could be the step needed for these technologies to become a ubiquitous teaching tool. The usage of middle-range priced hardware that has recently been specifically developed for this field might also be a necessary upgrade.

15. Acknowledgements

To Ana who made all this experience possible; to Pierre for the opportunity of working at his lab and the fantastic conditions; to Guillaume who guided me through the whole process and gave me access to his start-up resources and know-how; to Kshitij who was always ready to help in whatever necessary and would even send me huge eye-tracking analysis files through the internet; to Jessica who guided me through the experiment design; to Himanshu who sat right next to me and helped me with the most absurd questions; to Florence who manages the CHILI lab perfectly and guided me through all the bureaucratic procedures; to Francisco who guided me towards a proper dissertation.

16. References

Ainsworth, S. (1999). The functions of multiple representations. *Computers & Education, 33*(2), 131-152.

Association, A. P. (2001). *Publication manual of the American psychological association*: American Psychological Association Washington DC.

Ayres, P., & Cierniak, G. (2012). Split-Attention Effect *Encyclopedia of the Sciences of Learning* (pp. 3172-3175): Springer.

Dillenbourg, P., & Evans, M. (2011). Interactive tabletops in education. *International Journal of Computer-Supported Collaborative Learning, 6*(4), 491-514.

Fitzmaurice, G. W., Ishii, H., & Buxton, W. A. (1995). *Bricks: laying the foundations for graspable user interfaces.* Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems.

Gabrielli, S., Harris, E., Rogers, Y., Scaife, M., & Smith, H. (2001). *How many ways can you mix colour? Young children's explorations of mixed reality environments.* Paper presented at the CIRCUS 2001 Conference for Content Integrated Research in Creative User Systems.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain and Ireland*, 246-263.

Ishii, H., & Ullmer, B. (1997). *Tangible bits: towards seamless interfaces between people, bits and atoms.* Paper presented at the Proceedings of the ACM SIGCHI Conference on Human factors in computing systems.

Jordà, S. (2010). *The reactable: tangible and tabletop music performance.* Paper presented at the CHI'10 Extended Abstracts on Human Factors in Computing Systems.

Kaltenbrunner, M., & Bencina, R. (2007). *reacTIVision: a computer-vision framework for table-based tangible interaction.* Paper presented at the Proceedings of the 1st international conference on Tangible and embedded interaction.

Klemmer, S. R., Li, J., Lin, J., & Landay, J. A. (2004). *Papier-Mache: toolkit support for tangible input.* Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems.

Langolf, G. D., Chaffin, D. B., & Foulke, J. A. (1976). An investigation of Fitts' law using a wide range of movement amplitudes. *Journal of Motor Behavior, 8*(2), 113-128.

Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive science, 11*(1), 65-100.

MacKenzie, I. S. (1992). Fitts' law as a research and design tool in human-computer interaction. *Human-computer interaction, 7*(1), 91-139.

MacKenzie, I. S., & Buxton, W. (1992). *Extending Fitts' law to two-dimensional tasks.* Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems.

Marshall, P. (2007). *Do tangible interfaces enhance learning?* Paper presented at the Proceedings of the 1st international conference on Tangible and embedded interaction.

Price, S., Falcão, T. P., Sheridan, J. G., & Roussos, G. (2009). *The effect of representation location on interaction in a tangible learning environment.* Paper presented at the Proceedings of the 3rd International Conference on Tangible and Embedded Interaction.

Raffle, H. S., Parkes, A. J., & Ishii, H. (2004). *Topobo: a constructive assembly system with kinetic memory.* Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems.

Resnick, M. (1993). Behavior construction kits. *Communications of the ACM, 36*(7), 64-71.

Schneider, B., Jermann, P., Zufferey, G., & Dillenbourg, P. (2011). Benefits of a tangible interface for collaborative learning and interaction. *Learning Technologies, IEEE Transactions on, 4*(3), 222-232.

Shaer, O., & Hornecker, E. (2010). Tangible user interfaces: past, present, and future directions. *Foundations and Trends in Human-Computer Interaction, 3*(1–2), 1-137.

Zhang, J. (1997). The nature of external representations in problem solving. *Cognitive science, 21*(2), 179-217.

Zhang, J., & Norman, D. A. (1994). Representations in distributed cognitive tasks. *Cognitive science, 18*(1), 87-122.

Zufferey, G. (2010). *The Complementarity of Tangible and Paper Interfaces in Tabletop Environments for Collaborative Learning.* ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE.

Zufferey, G., Jermann, P., & Dillenbourg, P. (2008). A tabletop learning environment for logistics assistants: activating teachers. *Proceedings of IASTED-HCI 2008*, 37-42.

17. Appendix

Pre-Questionnaire

Age: _____ Gender: M □ F □

Study field: _____ Year of study: _____

1 Have you ever used an augmented interf ace? Yes □ No □

2 Have you ever used a tangible interface? Yes □ No □

3 I feel comfortable using these technologies:

Not comfortable	1	2	3	4	5	Very comfortable
	□	□	□	□	□	

4 Do you have some experience with warehouse organization? Yes □ No □

Intermediate Questionnaire (Lamp-Web)

5 I had no difficulty in understanding the system:

No difficulty	1	2	3	4	5	Very difficult
	□	□	□	□	□	

6 Using the system is intuitive:

Not intuitive	1	2	3	4	5	Very intuitive
	□	□	□	□	□	

7 My performance with the setup evolved along the time:

Disagree	1	2	3	4	5	Agree
	□	□	□	□	□	

Intermediate Questionnaire (Web-Lamp)

5 I had no difficulty in understanding the system:

No difficulty	1	2	3	4	5	Very difficult
	□	□	□	□	□	

6 Using the system is intuitive:

Not intuitive	1	2	3	4	5	Very intuitive
	□	□	□	□	□	

7 My performance with the setup evolved along the time:

Disagree	1	2	3	4	5	Agree
	□	□	□	□	□	

8 The video background was helpful:

Not helpful	1	2	3	4	5	Very helpful
	□	□	□	□	□	

9 Splitting attention between the screen and the tangibles had:

No impact	1	2	3	4	5	Big impact
	□	□	□	□	□	

Post-Questionnaire (Lamp-Web)

8 I had no difficulty in understanding the system:

No difficulty	1	2	3	4	5	Very difficult
	□	□	□	□	□	

9 Using the system is intuitive:

Not intuitive	1	2	3	4	5	Very intuitive
	□	□	□	□	□	

10 My performance with the setup evolved along the time:

Disagree	1	2	3	4	5	Agree
	□	□	□	□	□	

11 The video background was helpful:

Not helpful	1	2	3	4	5	Very helpful
	□	□	□	□	□	

12 Splitting attention between the screen and the tangibles had:

No impact	1	2	3	4	5	Big impact
	□	□	□	□	□	

13 The TinkerLamp was faster to use:

Disagree	1	2	3	4	5	Agree

	□	□	□	□	□	

14 The TinkerLamp was more accurate:

Disagree	1	2	3	4	5	Agree
	□	□	□	□	□	

15 The TinkerLamp was more intuitive to use than the TinkerWeb:

Disagree	1	2	3	4	5	Agree
	□	□	□	□	□	

16 The TinkerLamp was better to perform the warehouse task:

Disagree	1	2	3	4	5	Agree
	□	□	□	□	□	

17 I enjoyed more using the TinkerLamp:

Disagree	1	2	3	4	5	Agree
	□	□	□	□	□	

What did you think about the TinkerLamp/TinkerWeb?

_____

_____

_____

_____

_____

What are the main advantages of each device?

_____

_____

_____

_____

Post-Questionnaire (Web-Lamp)

10 I had no difficulty in understanding the system:

No difficulty	1	2	3	4	5	Very difficult
	□	□	□	□	□	

11 Using the system is intuitive:

Not intuitive	1	2	3	4	5	Very intuitive
	□	□	□	□	□	

12 My performance with the setup evolved along the time:

Disagree	1	2	3	4	5	Agree
	□	□	□	□	□	

13 The TinkerLamp was faster to use:

Disagree	1	2	3	4	5	Agree
	□	□	□	□	□	

14 The TinkerLamp was more accurate:

Disagree	1	2	3	4	5	Agree
	□	□	□	□	□	

15 The TinkerLamp was more intuitive to use than the TinkerWeb:

Disagree	1	2	3	4	5	Agree
	□	□	□	□	□	

16 The TinkerLamp was better to perform the warehouse task:

Disagree	1	2	3	4	5	Agree
	□	□	□	□	□	

17 I enjoyed more using the TinkerLamp:

Disagree	1	2	3	4	5	Agree
	□	□	□	□	□	

What did you think about the TinkerLamp/TinkerWeb?

_____

_____

_____

_____

_____

What are the main advantages of each device?

_____

_____

_____

_____