Tangible interaction on different technological setups

João Amaral Craveiro
Instituto Superior Técnico, joaocraveiro@tecnico.ulisboa.pt

Abstract **- This study investigates the differences and similarities between two tangible user interface setups in teaching 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 Web requires more adaptability and users clearly improve their performances over time. 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.**

Index Terms - Tangible User Interfaces, Education, Logistics, Technologic setups, User experiments

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 [1]. 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 [2]. According to [3] 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 [4]. 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 [5] 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.

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 CHILI lab at 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 [2].

To bring this prototype from the lab to the classrooms a startup called 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 [6]. 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.

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 *[7]*.

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 Massachusetts 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 [1]. 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 EPFL's 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 [8].

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 [9], 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 [10] 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 [11]. 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 [12]. 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.

TECHNOLOGY

The two setups considered for this experiment are the TinkerLamp and the TinkerWeb.

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 [13].

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.

To complement the setups and conduct the experiment other artifacts were used. Chili tags, 2D fiducial markers developed at EPFL allowed the detection and identification of all tangibles. A mobile eye-tracker was also used to trace the gaze of the subjects and allow a deeper analysis. Finally a powerful logging tool for the system was used to capture all events and supply the necessary data for the statistical analysis to be conducted.

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 [14]. 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 [15]. Appropriate representations can reduce the cognitive effort required by grouping information in the objects themselves [16]. 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 [17].

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 [18].

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 [19] but more recently these have been assessed as key factors on the learning experience [8] and some impacts have been identified experimentally [2]. 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.

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.

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.

TABLE I
EXPERIMENT CONDITIONS

TinkerLamp	TinkerWeb
Group1	Group2
Group2	Group1

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.

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.

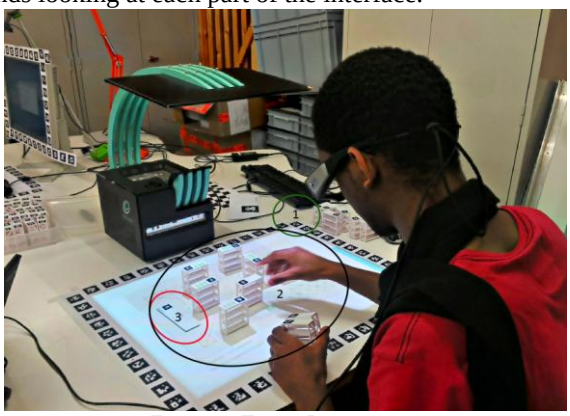

FIGURE 1: TINKERLAMP SETUP

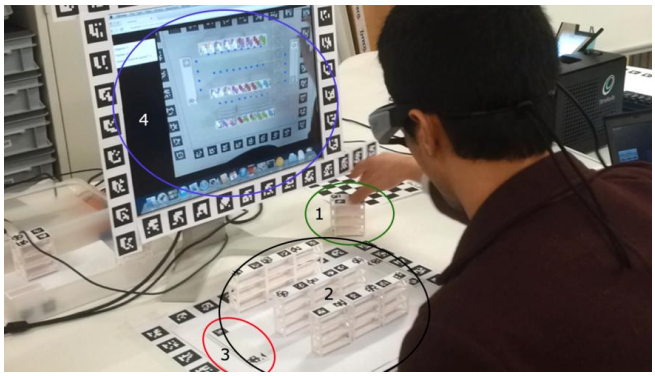

FIGURE 2: TINKERLAMP SETUP

The first set of tasks, interaction-only, 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.

TABLE II
TRIAL TYPES

Trial Type I 1 shelf	Trial Type II 2 shelves	Trial Type III 5 shelves	Trial Type IV 10 shelves
4 translations 4 translations w/ 90º rotation 4 translations 4 translations w/ 90º rotation	5 translations w/ 90º rotation	5 translations w/ 90º rotation 5 translations w/ 90º rotation and alignments	2 translations w/ 90º rotation and alignments

A cognitive task was also designed to evaluate a relevant dimension for the learning of the users. The tasks consist in

building a warehouse in 5 minutes. During this time the subjects must try to display 16 shelves in way that minimizes the average distance to both the in and the out docks, a typical real life problem. A pathway from the in to the out docks and accessibility of all shelves are other aspects they must account for.
All the tasks will be monitored using the log system. Because there are limitations to the score metric on the cognitive task the eye trackers will also be used to allow an expert evaluation. Two experts were asked to watch the performances and grade them from 0-5 in both strategy and optimization.

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. An example is provided as an appendix.
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.

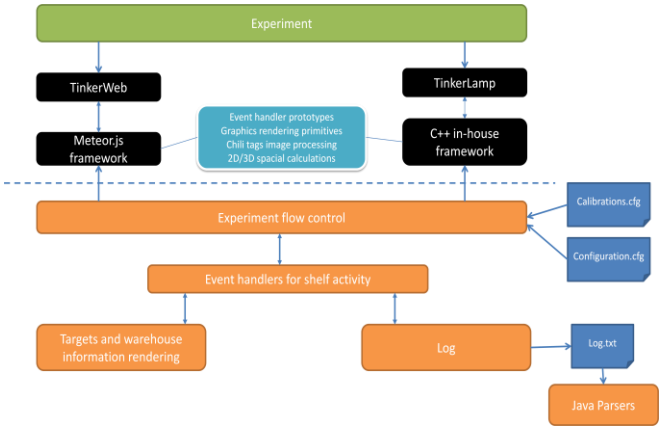

FIGURE 3: IMPLEMENTATION ARCHITECTURE

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 important tests during the 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 easily access 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.

EXPERIMENTAL RESULTS

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).

The first conducted analysis focuses on speed and 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.

TABLE III
SPEED ANALYSIS STATISTICS

	Between Subjects (n=5)	Within Subjects (n=10)
Bartlett test p-value	0.22	0.35
Shapiro-Wilk test p-value – Lamp	0.23	0.24
Shapiro-Wilk test p-value – 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]	[-291.23,2549.2]
Effect size	0.70	0.38
Power	0.13	0.19
Mean value (ms) – Lamp	1128.83	2296.69
Mean value (ms) - Web	1705.98	3425.69

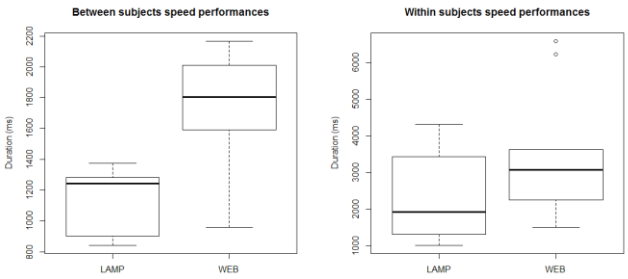

FIGURE 3

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.

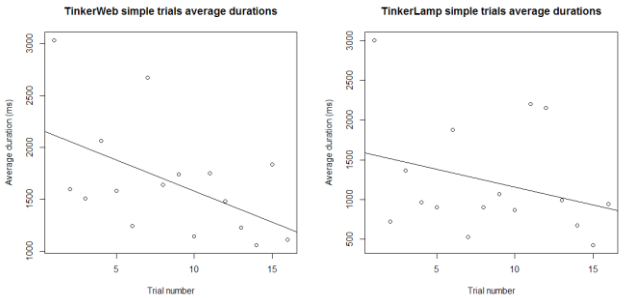

FIGURE 4

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.

The next step is to repeat the same procedure to accuracy. 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 III* 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%).

TABLE IV
ACCURACY ANALYSIS STATISTICS

	Between Subjects (n=5)	Within Subjects (n=10)
Bartlett test p-value	0.56	0.66
Shapiro-Wilk test p-value – Lamp	0-24	0.34
Shapiro-Wilk test p-value – 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]	[-2327.8,-786.9]
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

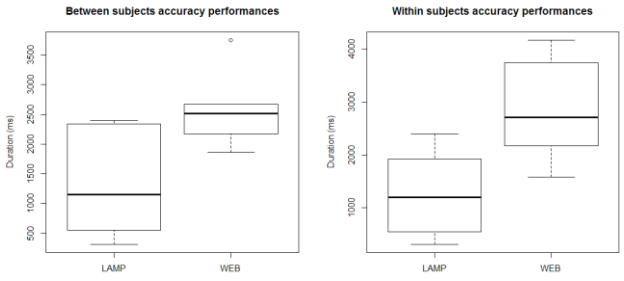

FIGURE 5

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.

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.

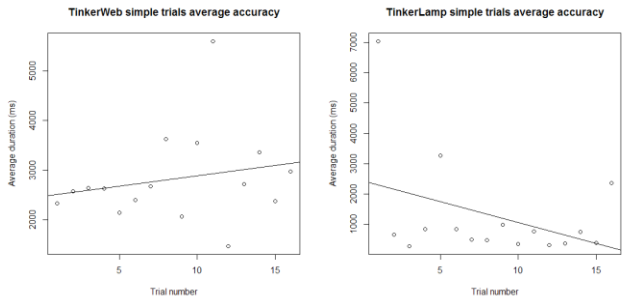

FIGURE 6

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.

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.

TABLE V
TRIALS DURATION STATISTICS

	Trial Type I	Trial Type II	Trial Type III	Trial Type IV
Bartlett test p-value	0.56	0.80	0.59	5.41e-07
Shapiro-Wilk p-value - Lamp	0.34	0.61	0.85	0.54
Shapiro-Wilk p-value – 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	-

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.

Concluded the interaction-only part of the experiment it's time to understand what happens cognitively.
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.

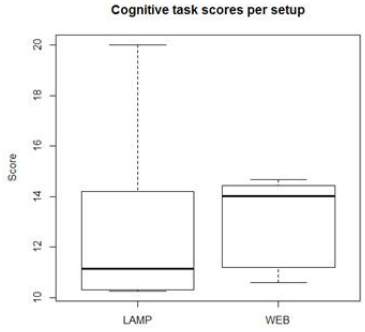

FIGURE 7

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).

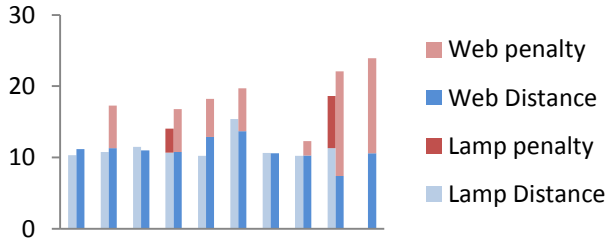

FIGURE 8: COGNITVE TASK SCORES

These results obtained represent the score metric, average distance to in and out docks, 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.
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.

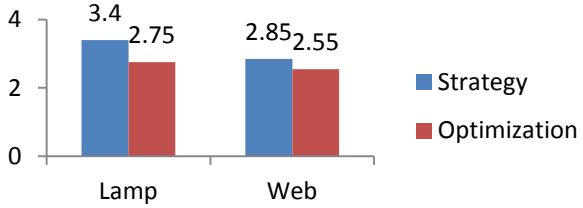

FIGURE 9: EXPERTS EVALUATIONS ON STRATEGY AND OPTIMIZATION

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.

TABLE VI
INTERACTION VS COGNITIVE TASK CORRELATION COEFFICIENTS

	Trial type I	All trials
TinkerLamp Strategy	-0.13	0.31
TinkerLamp Optimization	0.12	-0.10
TinkerWeb Strategy	-015	0.60
TinkerWeb Optimization	-0.45	0.62

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%.

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 [20].

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.

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.

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 coefficient of -0.43.

Comparing 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 one gets -0.64 for time on screen versus strategy score, -0.62 for optimization score and the opposite for the time on table, that is the complimentary of the time spent on screen. 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.

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.

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. Results show median of 0.8 seconds with statistical significance. It's important to note that the best performer on the web could perform as well as the best lamp performer, suggesting a bigger margin for familiarization with the setup.

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. The results were below statistical significance but still the median was an 1.3 seconds difference in favor of the TinkerLamp.

Trial duration confirmed what was 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.

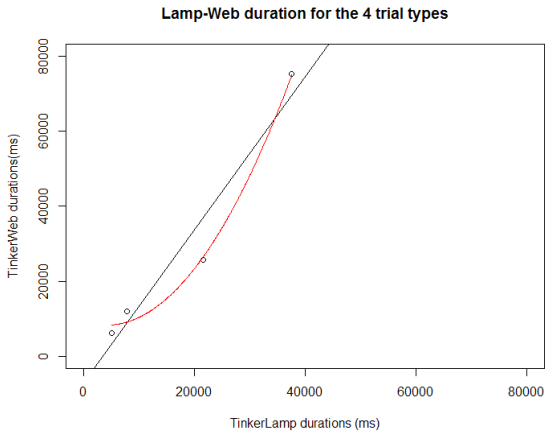

FIGURE 7: LINEAR AND QUADRATIC MODELS FOR COMPLEXITY IMPACT ON DURATION OF THE TRIALS

The 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. The web interface registered bigger penalties suggesting more difficulties leading with the warehouse as a whole. Again the best performances with the web attained the level of the lamp suggesting there is room for improvements and familiarization is very important.

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 lamp registered a 19% advantage in the classifications for strategy 11% for optimization. Again, the image of the warehouse as a whole, mirrored in strategy,

is more of a issue than optimization, mirrored somehow as the fine tuning necessary.

The split attention effect couldn't be significantly confirmed, although there was some evidence of it, opening the doors to something meaningful. 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.

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.

CONCLUSION AND FUTURE WORK

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.

REFERENCES

1. Ishii, H. and B. Ullmer. *Tangible bits: towards seamless interfaces between people, bits and atoms*. in *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*. 1997. ACM.
2. Schneider, B., et al., *Benefits of a tangible interface for collaborative learning and interaction.* Learning Technologies, IEEE Transactions on, 2011. **4**(3): p. 222-232.
3. Zufferey, G., *The Complementarity of Tangible and Paper Interfaces in Tabletop Environments for Collaborative Learning*. 2010, ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE.
4. Klemmer, S.R., et al. *Papier-Mache: toolkit support for tangible input*. in *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2004. ACM.
5. Shaer, O. and E. Hornecker, *Tangible user interfaces: past, present, and future directions.* Foundations and Trends in Human-Computer Interaction, 2010. **3**(1–2): p. 1-137.
6. Price, S., et al. *The effect of representation location on interaction in a tangible learning environment*. in *Proceedings of the 3rd International Conference on Tangible and Embedded Interaction*. 2009. ACM.
7. Fitzmaurice, G.W., H. Ishii, and W.A. Buxton. *Bricks: laying the foundations for graspable user interfaces*. in *Proceedings of the SIGCHI conference on Human factors in computing systems*. 1995. ACM Press/Addison-Wesley Publishing Co.
8. Dillenbourg, P. and M. Evans, *Interactive tabletops in education.* International Journal of Computer-Supported Collaborative Learning, 2011. **6**(4): p. 491-514.
9. Jordà, S. *The reactable: tangible and tabletop music performance*. in *CHI'10 Extended Abstracts on Human Factors in Computing Systems*. 2010. ACM.
10. Resnick, M., *Behavior construction kits.* Communications of the ACM, 1993. **36**(7): p. 64-71.
11. Raffle, H.S., A.J. Parkes, and H. Ishii. *Topobo: a constructive assembly system with kinetic memory*. in *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2004. ACM.
12. Kaltenbrunner, M. and R. Bencina. *reacTIVision: a computer-vision framework for table-based tangible interaction*. in *Proceedings of the 1st international conference on Tangible and embedded interaction*. 2007. ACM.
13. Zufferey, G., P. Jermann, and P. Dillenbourg, *A tabletop learning environment for logistics assistants: activating teachers.* Proceedings of IASTED-HCI 2008, 2008: p. 37-42.
14. Marshall, P. *Do tangible interfaces enhance learning?* in *Proceedings of the 1st international conference on Tangible and embedded interaction*. 2007. ACM.
15. Zhang, J., *The nature of external representations in problem solving.* Cognitive science, 1997. **21**(2): p. 179-217.
16. Larkin, J.H. and H.A. Simon, *Why a diagram is (sometimes) worth ten thousand words.* Cognitive science, 1987. **11**(1): p. 65-100.
17. Zhang, J. and D.A. Norman, *Representations in distributed cognitive tasks.* Cognitive science, 1994. **18**(1): p. 87-122.
18. Ainsworth, S., *The functions of multiple representations.* Computers & Education, 1999. **33**(2): p. 131-152.
19. Gabrielli, S., et al. *How many ways can you mix colour? Young children's explorations of mixed reality environments*. in *CIRCUS 2001 Conference for Content Integrated Research in Creative User Systems*. 2001.
20. Ayres, P. and G. Cierniak, *Split-Attention Effect*, in *Encyclopedia of the Sciences of Learning*. 2012, Springer. p. 3172-3175.