I. THE PROBLEM

Dominion Hosting Holding SpA (DHH) is a holding of companies in the Adriatic region of Europe providing website and application hosting, domain reselling, and other internet-related services. DHH.si Ltd is a subsidiary of DHH in Slovenia. Nearly half of all employees in these companies are helpdesk personnel, responsible for resolving issues reported by customers. These employees (Customer Care agents or simply agents) are users of the system we wanted to test.

We wanted to redesign the application used by Customer Service agents and see how the original application and the redesigned prototype compare to each other. Other than quantifiable usability metrics, we wanted to include metrics used to evaluate UX. The main research topic beside the difference in usability metrics (efficiency, efficacy, and satisfaction dimensions) is any difference in the emotional state of Customer Service agents after using the new prototype. On top of that, we also wanted to find out if Customer Service agents perceive the new application as aesthetically more pleasing and if they identified with the new prototype more than with the old application. We wanted to conduct these tests under laboratory settings although we wanted to replicate the working environment as closely as possible.

II. METHODOLOGY

While usability metrics in measurement procedures are well understood, documented, and standardized [4–6], user experience is less studied with a large number of different but less tested measurement tools.

We decided to use the definition of UX metrics as defined by [7] as the basis of our research and experiment. The authors recognized all the typical usability measures defined by aforementioned standards, but they grouped them in a different way. While ISO standards talk about three dimensions of usability (efficiency, effectiveness, and satisfaction), Tullis and Albert group metrics into four groups: performance, usability, self-assessment and physiological/behavioral metrics. Typical usability measures like time-on-task and number of clicks can all be counted into performance metrics. Success and error rate are part of the usability metrics. System Usability Score (SUS) is a typical self-assessment measure.

We also wanted to measure cognitive and physical effort. This can only be measured indirectly through other concrete quantitative measures. [7] emphasize that there is a correlation between number of “discrete actions” (e.g., clicks, key strokes …) and cognitive effort although we

Abstract—We conducted a study of a web-based helpdesk application used by Customer Care agents to provide information and support for customers. Ten Customer Care agents participated in a study on usability and user experience, where we compared the current application to a Prototype of a new version of the application. We used typical usability metrics to measure all three dimensions of usability (efficiency, effectiveness, and satisfaction). Furthermore AttrakDiff 2 was used to measure the hedonic and pragmatic quality of the product and attractiveness while Self-Assessment Manikin was used to measure affective reaction. The results show a statistically significant improvement or a positive trend in all usability measures. The results also show statistically significant improvement in hedonic stimulation quality and pragmatic quality measures for AttrakDiff 2 test. The Self-Assessment Manikin survey showed no significant change in affect between the two versions of the application.

INTRODUCTION

User experience (UX) is a relatively new field of research. It is rooted in human-computer interaction but has been both quickly recognized and criticized for being too vague by the same community [1]. The term was associated with various meanings and no single theory was developed within the design community [2]. Through time, it has developed from mainly programmatic research into a phase of conceptualizing the term and developing a research agenda [1]. There has been a lack of empirical research, which might be the reason that UX papers have not been frequently published in relevant journals.

This has changed in recent years as a number of empirical research papers were published by different authors including [3]. The authors note that there are several areas in which more work is needed. They describe four areas that need further research:

- Rich description of context,
- Changes in UX over a time period (for individual users and across a group of users),
- Multidimensionality, which forms the core of user experience,
- UX on desktop computers with productivity applications.

While usability has been thoroughly tested and standardized in different standards [4–6], user experience lacks in the aforementioned research. We chose to focus on UX in a desktop environment, which is still the most widespread type of use of computers and application in corporate environments.
should be careful if other metrics such as satisfaction, effectiveness and time measures counter these results [8].

To accurately measure user experience we must also include other dimensions such as perceived aesthetics and emotional response. For emotional response, there are at least three different methods to measure it: physiological reactivity, behavioural measure and affective reporting. Due to our limited resources, we decided to use affective reporting in a form of a self-reported survey of emotional state in a given moment. Different scales are used to measure different dimensions [9]. [10] introduced a Self-Assessment Manikin scale that was designed to be easily applicable to children and non-native English speakers without any loss of information. The scale measures three dimensions: arousal, dominance and happiness.

Aesthetics is another complex topic to measure but is an integral part of user experience. Although aesthetics was already a popular topic of research, a renewed interest was found in the end of the 20th century as a consequence of studying usability [11]. The concept of aesthetics is primarily linked to visual appearance but may also include other characteristics like sound, smell, or taste. It is a multidimensional concept as some studies identify separate properties that make up aesthetics such as complexity, ambiguity, novelty, and familiarity. [12] showed that a lower number of properties (typicality and novelty) are enough to describe product aesthetics. Lately, Lavie and Tractinsky [13] have developed a scale with two distinct categories named expressive and classical aesthetics. "Whereas classical aesthetics refers to design characteristics such as orderliness, clarity, and symmetry, design based on expressive aesthetics are rather original, creative, fascinating, and refined" [14]. [15] developed a tool called AttrakDiff 2 to measure four distinct dimensions (hedonic stimulation, hedonic identity, pragmatic quality and attractiveness) and showed that while hedonic quality and pragmatic quality are independent, attractiveness is determined by both factors.

III. THE EXPERIMENT

The goal of the experiment was to compare the helpdesk application that is currently used by Customer Service agents (Current System) with the prototype of the system designed on the feedback from the current system (the Prototype). We prepared a set of hypothesis we wanted to confirm (Table II.) or reject so we can estimate if a redesigned system really achieved the goal of reduced cognitive effort and better overall UX.

We conducted the experiment on ten individuals that form a representative sample. 50% of participants were female, 50% have high school education and the other half a university or Master’s degree. Area of education also differ, with participants ranging from general (gymnasium) through social (economics) to natural sciences (computer sciences, mathematics). Half the participants have been actively using the Current System for more than 2 years, while all participants use it at least occasionally. Although we used a small sample (10 people), this is actually the total population actively using the system. Furthermore, the experiment was designed to focus on experienced users. The reasoning behind this is that the system was designed for knowledgeable users and as such must facilitate quick information retrieval even at the expense of a steep learning curve for new users.

We conducted the experiment in a laboratory setting with a desktop personal computer in a dual monitor setup. Participants typically use this setup in a work environment. The workplace was also equipped with headphones (regularly used for telephone communication with the customers) and a web camera to capture the face of participants. The session was filmed (combining video of the participant, the desktop of the computer, and audio communication with the participant)

We gathered all the participants at the beginning of the experiment. We explained the first phase of the experiment without revealing its goal and dual-phase nature. Customer Care agents are accustomed to continuous education that covers new features of the system, but they perceived this experiment as an exam of their capabilities. We explained the nature of user experience testing, that we are measuring their system and not them individually. We also explained that we are not interested in performance and the metrics of particular participants but in the statistical significance that is observed over the whole population. We explained what data will be gathered and how it will be processed and defined the rules of communication (how to declare that a task was finished, that they should be role playing and similar).

When the experimenter started conducting the first phase of the experiment, he made a quick recap of what we already explained to each individual. Each participant was given 5 minutes to set up the computer (install browser add-ons, open all the necessary windows and enter all the passwords, so they are remembered by the application, etc.). After this, participants were isolated in a room and communication conducted only via phone.

We started the experiment by making a phone call to the participant. Participants used “soft phones” (a desktop application that they usually use to communicate with the customer – 3CX Phone). We then began filming the experiment. The experimenter first introduced a General Survey, followed by a Self-Assessment Manikin (SAM) survey. The experimenter encouraged the participant to select images from the SAM questionnaire that they identified with.

The experimenter then informed the participant that they would begin the tasks. The tasks for phase 1 of the experiment are listed in Table I. Each task was a role-play between experimenter and the participant. This means that the experimenter acted as a customer and requested

<table>
<thead>
<tr>
<th>#</th>
<th>Hypothesis description</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td>Prototype usage improved Customer Care agents' success rate.</td>
</tr>
<tr>
<td>H2</td>
<td>Prototype usage increased Customer Care agents' time spent on tasks.</td>
</tr>
<tr>
<td>H3</td>
<td>Prototype usage increased Customer Care agents' number of mouse clicks.</td>
</tr>
<tr>
<td>H4</td>
<td>Prototype usage increased Customer Care agents' number of mouse movements.</td>
</tr>
<tr>
<td>H5</td>
<td>Prototype usage reduced Customer Care agents' error rate.</td>
</tr>
<tr>
<td>H6</td>
<td>Prototype usage improves Customer Care agents' affective experience.</td>
</tr>
<tr>
<td>H7</td>
<td>Prototype usage improves “subjective perception” of user experience.</td>
</tr>
</tbody>
</table>
After participant completed all the tasks, the experimenter presented the last three surveys, beginning with the second SAM survey, continued with standardized SUS survey and finished with AttrakDiff survey.

This procedure was repeated again after we developed the Prototype. An analogous set of tasks was used due to small sample size while all other criteria remained the same as in the first phase.

IV. RESULTS

In our experiment, we measured agent’s success rate in Phase 1 (first measurement using the Current System) and in Phase 2. This was done to accept or reject H1 (Prototype usage improved Customer Care agents’ success rate.). Each measurement allowed for only two outcomes: successful and unsuccessful. The average success rate of first measurement was $M = 81.11$, $SD = 20.98$. In the second measurement the average success rate was higher $M = 92.22$, $SD = 7.50$. We used nonparametric test (Wilcoxon signed-rank test) to investigate if the difference in success rate was statistically significant. We assumed that the distribution of success rates was symmetrical. Descriptive data shows a consistent trend of improvement. A Wilcoxon signed-rank test determined that there was a statistically insignificant increase of success rate when subjects used the Prototype compared to the no Prototype use $z = -1.826$, $p = .068$. Since the $p$-value is relatively close to the borderline (.025), we can therefore assume that there is a positive trend, which led us to decide to report the results. The improvement of success rate was statistically insignificant which means that we cannot confirm nor reject the null hypothesis of the H1.

Eight out of ten agents spent less time solving tasks while testing H2 (Prototype usage increased Customer Care agents’ time spent on tasks). The average time spent on solving nine tasks in Phase 1 was $M = 829.07$ s, $SD = 53.20$s and $M = 642.12$s, $SD = 49.41$s in the second measurement. When using the Prototype, agents spent 186.95 seconds ($SE = 68.98$) less on tasks. With a paired samples t-test, we determined that there was a statistically significant decrease of overall time spent on task when subjects used the Prototype compared to the Current System $t(9) = -2.97$, $p = .016$, $d = 0.94$.

There was a statistically significant difference between means ($p < .025$), and therefore, we can reject the null hypothesis and accept the alternative hypothesis H2.

Nine out of ten agents performed less mouse clicks while solving tasks (test for H3 - Prototype usage increased Customer Care agents’ number of mouse clicks). The average number of mouse clicks agents made while solving nine tasks in the first measurement was $M = 171.20$, $SD = 18.52$ and $M = 64.40$, $SD = 11.15$ in the second measurement. When using the Prototype, agents made 106.80 fewer mouse clicks ($SE = 24.03$) on tasks. With a paired samples t-test we determined that there was a statistically significant decrease in mouse clicks when subjects used the Prototype compared to the Current System $t(9) = -4.44$, $p = .002$, $d = 1.40$.

There was a statistically significant difference between means ($p < .025$), and therefore, I can reject the null hypothesis and accept the alternative hypothesis H3.

<table>
<thead>
<tr>
<th>#</th>
<th>Task Scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Could you please tell me if you received payment for order #1551240?</td>
</tr>
<tr>
<td>2</td>
<td>Please, place an order in my name for package HiskaPlus, ID #45929.</td>
</tr>
<tr>
<td>3</td>
<td>I am calling from XYZ, please tell me if I have any outstanding unpaid invoices and the amount of credit available to me?</td>
</tr>
<tr>
<td>4</td>
<td>I did not receive a notification about the expiration of domain XYZ.si. When did you last notify me?</td>
</tr>
<tr>
<td>5</td>
<td>I would like to redirect the domain XYZ.com to another IP address. How do I do that?</td>
</tr>
<tr>
<td>6</td>
<td>I would like to change the content of my web site. XYZ told me you are responsible. I am XY from X and the website is on domain Zsi</td>
</tr>
<tr>
<td>7</td>
<td>I would like to create a new e-mail account for domain XYZ.si. Please guide me!</td>
</tr>
<tr>
<td>8</td>
<td>I am calling from XYZ. I paid for a domain Y.si but did not receive an invoice. Please send it to <a href="mailto:Z@X.com">Z@X.com</a></td>
</tr>
<tr>
<td>9</td>
<td>I am calling from XYZ I received a notification that my hosting will expire. Could you please check if I have any active domains in my package and if I could reduce the cost of subscription?</td>
</tr>
</tbody>
</table>
All agents made fewer mouse movements while solving tasks (test for H4 - Prototype usage increased Customer Care agents' number of mouse movements.). The average number of mouse movements agents made while solving nine tasks in the first measurement was $M = 191,577.62$, $SD = 56,111.84$ and $M = 109,834.27$. In the second measurement. When using Prototype agents made $81,743.55$ mouse moves ($SE = 18,898.60$) less on tasks. With a paired samples t-test I determined that there was a statistically significant decrease of mouse movements when subjects used the Prototype compared to the Current System $t(9) = -4.33$, $p = .002$, $d = 1.37$.

There was a statistically significant difference between means ($p < .025$), and therefore, I can reject the null hypothesis and accept the alternative hypothesis H4.

Six agents decreased their error rate in the second measurement of testing for H5 (Prototype usage reduced Customer Care agents' error rate). The average error rate of first measurement was $M = 1.50$ $SD = 1.50$. In the second measurement the average error rate was lower $M = 1.10$, $SD = 1.45$. To investigate if the difference in error rate was statistically significant, we used nonparametric test (Wilcoxon signed-rank test) and assumed that the distribution of error rates was symmetrical. A Wilcoxon signed-rank test determined that there was a statistically insignificant decrease of error rate when subjects used the Prototype compared to the Current System $p > .025$.

Due to the statistically insignificant change in error rate, we can neither confirm nor reject the null hypothesis H5.

We tested hypothesis H6 (Prototype usage improves Customer Care agents' affective experience) with the use of Self-Assessment Manikin. Affective experience with this tool is measured with three dimensions – arousal, dominance and pleasure. The average level of all three dimensions slightly rose. In the first measurement average arousal was 2.33, dominance 3.33 and pleasure 3.68. In the second measurement average arousal level was 2.60, dominance 2.60, and pleasure 4.00.

All changes were positive, but statistically insignificant which means that I can neither confirm nor reject the null hypothesis H6.

I investigated four aspects of subjective perception of the quality of user experience while testing for H7 (Prototype usage improves “subjective perception” of user experience.) – pragmatic quality (PQ), hedonic quality (HQ-I, HQS) and attractiveness (ATT). Average levels of all four dimension were higher when using the Prototype.

Average PQ level rose from $M = 29.30$, $SD = 1.95$ to $M = 34.10$, $SD = 4.36$. The increase of average level of pragmatic quality 4.80 ($SE = 1.79$) was statistically significant $t(9) = 2.68$, $p = .024$, $d = 0.85$.

Average HQI level rose from $M = 29.00$, $SD = 4.19$ to $M = 35.80$, $SD = 5.43$. The increase of average level of HQI $6.80$ ($SE = 2.72$) was statistically insignificant $t(9) = 2.50$, $p > .025$, but close to the borderline ($p = .034$).

Average HQS level rose from $M = 27.40$, $SD = 1.73$ to $M = 3.40$, $SD = 1.41$. The increase of average level of HQS 8.00 ($SE = 2.34$) was statistically significant $t(9) = 3.42$, $p = .008$, $d = 1.08$.

Average ATT level rose from $M = 30.70$, $SD = 1.27$ to $M = 36.70$, $SD = 1.73$. The increase of average level of ATT 6.00 ($SE = 2.66$) was statistically insignificant $t(9) = 2.25$, $p > .025$.

Changes in the attractiveness (ATT) and hedonic quality (identity) (HQ-I) scales were statistically insignificant. Changes on hedonic quality (stimulation) (HQ-S) and pragmatic quality (PQ) showed significant improvement between two phases which means I can accept the null hypothesis H7.

V. DISCUSSION AND FURTHER WORK

Comparison between two phases showed an overall improvement in all parameters for both individual subjects and as an average over the whole group. Some of the increases are statistically insignificant due to the small population sample. We measured usability on all three dimensions: efficiency, effectiveness, and satisfaction. All measures showed a statistically significant improvement or insignificant change. The average SUS score rose 15 points between phases from 60 to 75 and is statistically significant. The score is above industry average [7, 17]. We believe that the increasing confidence throughout the study is also one of the reasons for a better overall SUS score achieved with the Prototype. In total, agents used 22% less time on tasks, 61% fewer mouse clicks and moved the mouse 42% less than when working with the Current System. I believe that this finding clearly shows that we made correct design decisions based on analysis of the first phase of the experiment. This result is also the greatest success of the study. The central challenge of this thesis is to design a system that would reduce the cognitive effort required by the Customer Care agents, thus freeing them up to focus on customers’ issues. As the measure of cognitive effort is by definition linked to user interactions and time on task [7], we can simply rely on efficiency measures to provide an estimate of cognitive effort. These results clearly show that the features of the Prototype are exactly the features that played a major role in decreasing the cognitive effort.

Self-Assessment Manikin tests showed insignificant changes in agents’ mood. Even though tests were performed before and after all the tasks through both phases, emotional state on all three dimensions remained similar. I believe that this is due to a small sample size. When analyzing the AttrakDiff 2 scale, we were interested in pragmatic quality of the product, hedonic stimulation and attractiveness factors. In my case, we see a statistically significant rise in average values when comparing Current System to the Prototype. This finding is in line with the effectiveness and efficiency usability measures. Thus, I can conclude that agents found the Prototype easier to use than the Current System. I found a statistically insignificant increase in the Attractiveness dimension between two phases. I also found that pragmatic and hedonic qualities have an impact on attractiveness, which is also in line with other studies like [18].

We plan to further develop the Current System and upgrade it with tools that were implemented in the Prototype. We also plan to conduct additional experiments when we complete these features to determine if changes have the same effect as in this study. We also plan to implement features that were requested by agents when testing the Prototype (additional options for dig tool, options to hide expired services, etc.). The next step in
developing this product is the implementation of some kind of natural language-processing text-based interface. This would allow the agent to search for information and conduct certain tasks by writing commands in a natural language.

REFERENCES