Evaluating Design Variations using UX-Analyzer

Juan Cruz Gardey
LIFIA, Fac. de Informática, Univ.Nac. de La Plata
La Plata, Buenos Aires, Argentina
jcgardey@lifia.info.unlp.edu.ar

Julián Grigera
LIFIA, Fac. de Informática, Univ.Nac. de La Plata
CONICET
CICPBA
La Plata, Buenos Aires, Argentina
juliang@lifia.info.unlp.edu.ar

Andrés Rodriguez
LIFIA, Fac. de Informática, Univ.Nac. de La Plata
La Plata, Buenos Aires, Argentina
arodrig@lifia.info.unlp.edu.ar

Alejandra Garrido
LIFIA, Fac. de Informática, Univ.Nac. de La Plata
CONICET
La Plata, Buenos Aires, Argentina
garrido@lifia.info.unlp.edu.ar

ABSTRACT

User Experience (UX) is an essential factor in today's business, and so large companies have adopted a process of continuous UX monitoring. In this process, design experts run user tests and analyze UX performance indicators, which are usually obtained from standardized questionnaires performed with users. Thus, this process is expensive, as it requires paying test participants, and allocating the time to manually analyze the collected data. Previously, we have defined interaction effort as a metric for the dynamic performance of web elements, based on measures of the user interaction, and we presented UX-Analyzer as a tool to visualize the interaction effort. UX-Analyzer allows UX experts or other team members to evaluate web pages automatically and in a transparent way. Visualizing the interaction effort may provide a good indication towards the level of UX of an online system. It may also be used in the context of an A/B testing approach that instead of revenue or conversions, it compares the UX of alternative versions of a website. In this paper we show how UX-Analyzer can perform in real settings. We present a case study with 152 participants that show its applicability, and we also add a qualitative study with design professionals to assess its adoption and usefulness. Additionally, we have extended the related work, and describe some new features that improve the tool's flexibility.

CCS CONCEPTS

• **Human-centered computing** → **Visualization systems and tools**; **HCI design and evaluation methods**;

KEYWORDS

User Interaction, User Experience, A/B testing, Machine Learning

1 INTRODUCTION

User Experience (UX) is crucial for the success of any software product, particularly for online systems that are exposed to a large number of users [2, 31, 46]. The term UX not only refers to the objective dimensions of user interaction with a product or service, such as efficiency or effectiveness, but also includes subjective aspects such as emotions and comfort [25].

Evaluation is a key task in building high-quality UX [28]. Although the investment of resources required for evaluation is recognized by companies, these UX practices are often neglected due to time constraints and the consequent impact on product quality. These constraints pose challenges for small and medium-sized agile teams working in short development cycles [1]. In this context, our motivation is to propose automated solutions for agile teams working under time pressure and with scarce resources to evaluate the UX of their products. We seek solutions that reduce the cost of evaluating and improving UX throughout the product life cycle, particularly in the context of processes with frequent deliveries.

A usual practice of large companies to periodically measure their level of revenue or conversions is to apply automatic controlled experiments like A/B testing [43]. In these experiments, each user is randomly exposed to one of different variants of a design, and a precise metric is used to select the outperforming variant [26]. However, the metric used rarely includes usability or UX [43].

In our previous work, we developed a metric called *interaction effort* to automatically evaluate and compare alternative designs [20, 21]. This metric is used to rate user behavior around individual widgets, and can be predicted using low level interaction logs as the sole input. Since interaction effort is calculated on single user interface (UI) elements, it facilitates localizing user interaction issues, but it can also be used for comparing small design variations. A key benefit of this metric is that it is transparent to the users and easy to calculate.

In a recent study we showed another use for the interaction effort metric: to provide a single effort score for a complete UI, by aggregating the scores of its widgets [19]. To make this available to UX experts, we created UX-Analyzer: a tool to calculate interaction effort and compare alternative versions of a specific web page [18]. In this way, UX-Analyzer shows different aggregations of the interaction effort score that give the UX expert varied perspectives of the effort required by the analyzed UI. The tool captures interaction logs from user sessions, which are fed into a set of prediction models to get the interaction effort on each widget contained on

 This work is based on an earlier work: UX-Analyzer: Visualizing Interaction Effort for Web Analytics, in SAC'24 Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing, pp.1774–1780, ACM, 2024. http://dx.doi.org/10.1145/3605098.3636013.

the target page. Then, the resulting scores are aggregated to compose the global effort of the UI under analysis. Again, the tool is transparent to users, since it only requires embedding a script in the target page.

UX-Analyzer is directed at UX experts, who can use it to evaluate different designs with a few users to validate design changes, or even with larger masses of users in online experiments like A/B testing. Although A/B testing is a suitable approach to evaluate UX, most existing A/B testing tools focus on measuring a company's revenue with conversion rates, which are not a direct indicator of UX [38]. Instead, with the interaction effort the evaluation of alternative designs can focus specifically on UX.

In this work we extend the findings of our latest work on UX-Analyzer [18], by evaluating the tool's performance in more realistic settings. With this objective we conducted 2 studies: a new case study with 152 participants that showcases it capabilities, and a qualitative analysis of its features, adoption potential and usability through interviews with UX professionals. We also discuss pros and cons of different approaches for automatic UX evaluation in the related work. Additionally, we describe some new features added to UX-Analyzer.

2 BACKGROUND AND RELATED WORK

User experience is a key component in the success of an interactive product. The ISO standard defines UX as "user's perceptions and responses that result from the use and/or anticipated use of a system, product or service" [25]. According to this definition, UX is a broad concept that includes both instrumental factors related to the user's performance while interacting with a system (effectiveness, efficiency) and hedonic aspects that are subjective to the users, such as aesthetics, enjoyment, comfort, and pleasure. Therefore, UX emerges as a consequence of the user's internal state (expectations, needs, etc.), characteristics of the system (complexity and functionality), and the context in which the interaction occurs. [23].

Although the first release of a product should have an acceptable UX level, it is of utmost importance to evaluate UX in further development cycles [8, 41]. Several studies have investigated the difficulties of synchronizing agile development cycles with UX practices, such as a lack of time and resources and deficient team communication [12]. Consequently, only large companies can afford manual evaluation methods that typically involve user testing at the end of each iteration [16, 44].

An established manual method for quantifying the UX level from one iteration to the next is to use questionnaires [24]. Using this method, after a round of user tests, the participants answer different statements with values from a typical Likert scale. Some well-known questionnaires include the User Experience Questionnaire (UEQ) [27], Standardized User Experience Percentile Rank Questionnaire (SUPR-Q) [42] and UEQ+ framework to create questionnaires from different UX scales [34]. Although questionnaires efficiently measure a user's subjective attitude towards the system being evaluated, any manual method is expensive because it requires recruiting participants, and the results can be biased because of the staged setting [37].

Some studies have focused on automating this process to provide affordable solutions for UX evaluation. Bakaev et al. classified automatic evaluation methods into three classes: metric-based (help UX experts during inspection methods on static webpages), interaction-based (log real user interaction during user tests), and model-based (users and their interactions are simulated to create and train models) [3].

Metric evaluation methods help experts to check guidelines [6] and assess UX factors that can be analyzed from a static perspective, such as the aesthetics of a website [13, 35, 45] or visual design [14]. Guideliner [32] assesses web UI conformance to usability guidelines during the development phase. It contained a predefined repository of usability guidelines and concrete assessment metrics. The tool Guideliner analyzes the UI code, extracts information about UI elements and their features, and compares it to the metrics defined in the usability guidelines to determine whether the user interface meets the guidelines [32]. A drawback of metric-based approaches is that it remains difficult to determine the significance of metrics for different user tasks and contexts of use [3].

Interaction-based evaluation methods capture and analyze event logs during remote user testing or in a real context, and provide tools to visualize problematic interactions [40] or detect potential usability problems [22]. Tools that show heat maps of clicks, such as ClickHeat [15] and eye-tracking visualizations [5] may also be considered in this category. The drawbacks of interaction-based methods are that they may be difficult and time-consuming to interpret because they require a large number of samples [3], do not provide a unified score for a design that can serve as a basis for subsequent development cycles, such as questionnaires do [36], and/or require specialized hardware or an in-lab setup.

Finally, model-based evaluation methods predict user interactions, such as the well-known KLM [9] or Fitt's Law [17]. The drawback of these two models is that they capture one dimension of user interaction in isolation, making it difficult to predict realistic interaction tasks [30]. For instance, Bertram and Dahm describe the design and implementation of a Figma plugin that performs automated usability assessments on UI prototypes. It combines metric and model-based approaches with usability metrics derived from standards, such as Quality in Use Integrated Map (QUIM) [39] and the KLM model [9] to predict interaction times and identify patterns that suggest usability issues. The plugin allows interaction sequences to be defined, evaluated, and compared, thereby providing visualized results through graphs and reports [4]. Some limitations of using this model-based evaluation are the level of abstraction required by users and the difficulty in understanding the task definition process with KLM.

Another example of a mixed approach is the WaPPU tool, which combines interaction and model-based evaluations. It has machine learning models that predict various usability aspects of web pages (confusion, distraction, readability, etc.) from user interaction records and user surveys [43]. However, the main limitation of WaPPU is that the interaction logs used as features for prediction are highly coupled to the structure of the target page; therefore, new models must be developed every time a different page has to be evaluated.

Our proposal is also a mix of interaction and model-based evaluations, such as WaPPU, as it involves logging user interaction

events and using these data to build models that predict user behavior [18]. In our case, the models predict a single score called the Interaction Effort metric [20]. This metric was inspired by the concept of cognitive load (CL), which describes a user's mental effort required by a system to interact with it [10]. Although some of this CL is intrinsic to the task at hand, suboptimal design can also lead to greater cognitive load [11]. Thus, interaction effort indicates the interaction cost related to the physical and mental effort that users make to achieve their objectives with a system [7]. In the context of e-learning environments, others studies have measured workload in relation to the interaction behavior, specifically with video lectures [33], and the attention level of students as measured with webcam and mouse behavior [29].

3 UX-ANALYZER IN ACTION

UX-Analyzer is a tool that captures several micro-measures from real interactions with a web application, and using prediction models allows visualizing and manipulating the interaction effort at different levels. The tool was created mainly for UX experts, who can analyze the resulting interaction effort and propose design changes in response to what they observe. In addition, UX-Analyzer was designed to be intuitive enough to be used by others team members such managers or product owners, who may be interested in monitoring the UX of a system.

UX-Anayzer is web application in which the users (designers and other stakeholders) can sign up and create evaluations to observe different pages of a target application. An evaluation contains one or more versions of the observed pages.

The tool also includes a JavaScript code that must be pasted in the target application to capture the user sessions of a specific version and send them to the server. Each user session contains the interaction logs that are fed into the prediction models to estimate the interaction effort of each widget in the analyzed page. These scores are then used to calculate the global effort of the version.

In the following subsections, we first describe the original concept of interaction effort for individual widgets, and how UX-Analyzer aggregates the score of widgets into sessions and webpages. Next we describe the main concepts that UX-Analyzer manipulates: evaluations, versions, user sessions, and widgets.

3.1 Interaction Effort

Interaction effort is a score that represents if and how much users struggle with individual webpage elements [20]. This happens when a widget is not the most appropriate for the task, or input required, causing discomfort and demanding extra time and effort from users. We built prediction models for 6 different types of widgets: text inputs, selection, link/buttons, radio-buttons, date pickers and date selects. Depending on the type of widget, several micro-measures are captured from real interactions, such as the time spent on a specific widget, mouse movements, keystrokes, among others [20]. The models were created by having UX experts observe widgets' interaction and assign a score that ranges from 1 to 4 according to their subjective analysis. This scale allows to score an interaction as effortless (1 and 2) or demanding (3 and 4).

The main goal of having a widget-centered metric is to evaluate bounded portions of a UI. This may occur when alternative versions of a design are proposed with small variations among versions, as it usually happens in an A/B testing setting. However, even small variations may involve more than one widget, variations with drastic changes may be desired, or it may be necessary to run several experiments with different users and contexts before reaching a decision. Thus, by combining the interaction effort of multiple users and widgets, the metric can also give UX experts a broader perspective of the target UI.

UX-Analyzer may be used to calculate and visualize the global interaction effort of a webpage, as well as the aggregations by user and by widget to provide a full picture of the user interaction effort. Figure 1 shows how the global interaction effort is calculated. Besides the global score, the interaction effort of each user with each widget can also be aggregated either by user or by widget. These aggregations can provide additional feedback about the user experience of a UI, for instance to identify specific users that make high effort or to detect certain widgets that are problematic for many users.

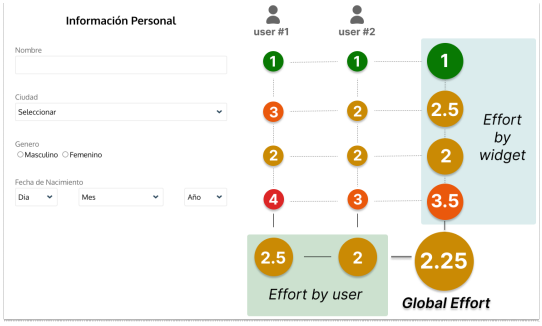

Figure 1: Interaction effort aggregated by users and by widgets.

3.2 Website's Evaluations

A UX expert interested in analyzing alternative designs of a target application should first create an *evaluation* in UX-Analyzer. Since the target application may have multiple pages with a wide range of functionalities, the purpose of the evaluation is to group the different pages and their versions to be analyzed. The evaluation size will depend on what the user needs to evaluate. For instance, an evaluation could assess general aspects such as the checkout process on a e-commerce, as well as more detailed concerns like a specific set of fields in a form.

When a designer logs in UX-Analyzer, the first view is the evaluations' list they have created. Each evaluation is identified by a name that describes the design aspect being analyzed.

3.3 Versions

The application versions are design variants that belong to an evaluation. A version can also be defined as a set of user sessions collected on specific pages that are used to estimate the interaction effort. When a new version is created, UX-Analyzer generates a code snippet to include in the target application for capturing the user interactions with the widgets. This snippet contains a specific

version token that is used by the backend of the tool to determine the target version of each user session received.

To create a new version, the designer must provide a name and the URLs of the pages that will be included in it. Although the global interaction effort was thought for individual web pages, in some cases it may be necessary to know the effort score for a group of related pages. In this way, a version can contain one or more URLs, which then will be used to filter the user interaction logs on the selected pages when a new user session arrives to the server.

The main view of the evaluation displays a list with all the generated versions (Figure 2). There, the expert can observe for each version, the global effort score and the amount of user sessions captured so far. This score is updated as new sessions are sent to the server. The expert can also click on a version to visualize its details, like the user sessions and the widgets included in the target pages.

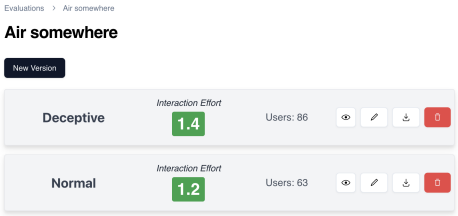

Figure 2: Versions list of an evaluation.

3.4 User Sessions

The code snippet generated by UX-Analyzer when creating a version is used to capture the user interaction logs. These logs are basically the micro-measures that result from the user interaction with each widget under analysis. The interaction logging starts when a user navigates to a URL included in the version, and it ends when the page is abandoned. For each widget, besides the micro-measures the script also logs the URL and its XPath, which then together allow to identify that widget instance. When the logging ends, the session duration is calculated and all the information is sent to the backend. There, UX-Analyzer processes the logs and store them in the version identified by the received token.

Figure 3 illustrates how the captured user sessions look in UX-Analyzer. The most relevant attribute of each user session is the interaction effort that results from aggregating the predicted effort on all the widgets the user interacted with. This aggregated score is the *effort by user* shown in Figure 1. In addition to the effort score, the sessions also show their date and duration.

This report is useful to monitor the users that interact with a particular version. By analyzing the effort score of each session and its duration, it is possible to detect particular cases in which a user experiments interaction problems.

3.5 Widgets

In the context of a version, the tool also allows to visualize the widgets list included in the target pages (Figure 4). Given that each widget is identified by its XPath and the URL, UX-Analyzer obtains this list grouping the interaction logs by these attributes. Each

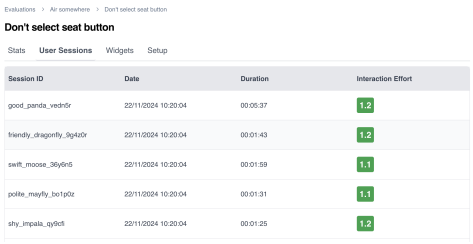

Figure 3: User sessions of a specific version.

widget contains the averaged interaction effort (*effort by widget* in Figure 1), and a label along with its type (text input, link, etc.) to recognize it. Moreover, the widgets have an associated weight that determines their influence in the global effort score of the underlying version. According to the intended analysis, the weights may be adjusted to give more importance to specific widgets and these changes are reflected in the global score. Moreover, widgets can also be disabled so they are not considered at all when obtaining the overall score. In this way, UX experts can center the analysis on the widgets that are crucial for the success of the target website, and leave those less important widgets out or in a second place.

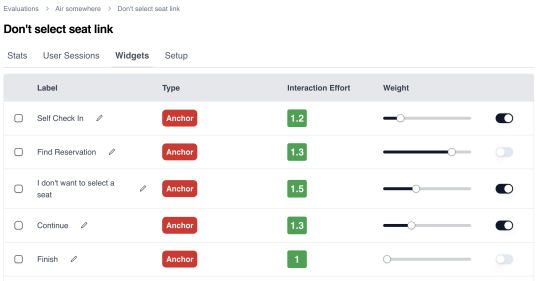

Figure 4: List of widgets that users interacted with using a specific version.

Like the effort of user sessions, the interaction effort aggregated by widget is also important to understand the value of the version global effort. In particular, this aggregation allows to detect the UI elements that users are struggling with.

4 USE CASE

In order to assess the applicability of the tool, we used UX-Analyzer to evaluate and compare two alternative designs for a specific website. This website is an airline online system in which users can complete a flight check-in, make a reservation, check flights status, etc. We developed a dummy version of the website in order to capture interaction data and avoid sending sensitive information to a real system. We also changed the website look and feel to prevent user bias of prior experiences with a specific airline or brand. We called the website “Air somewhere". (see Figure 5).

We focused the evaluation on the check-in flow, in which the user enters a reservation code, validates the personal information, and finally has the chance to select a specific seat for an additional fee. We specifically intended to analyze the effort of the seat selection

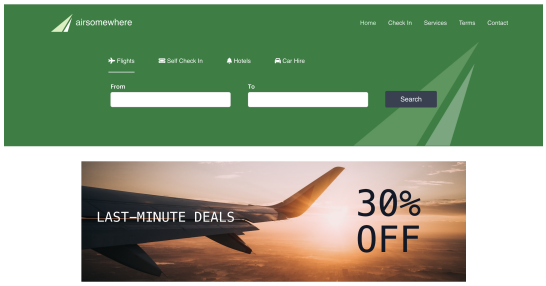

Figure 5: Website evaluated with UX-Analyzer.

step using two versions that are commonly seen on the web. Thus, we created two alternative designs for the seat selection page, which are shown in Figure 6.

In both designs, the seat selection page shows an airplane's seat map with all the available seats and their different costs, depending on their location. Users have the choice of selecting a seat for extra money, which enables the "Continue" button, or selecting the "I don't want to select a seat" option. The element for selecting this option is what changes between the alternative designs: the *Deceptive* version shows the option as a link, with its text noticeably smaller than "Continue". This decision is intended to make users focus on the "Continue" option, leading them to pay for a seat. This kind of designs that trick users to get a benefit are often called "Deceptive Patterns", and they are very common nowadays on the web. On the other hand, the *Normal* version implements a button in which the caption "I don't want to select a seat" has the same visual hierarchy as "Continue". In this case it is more evident for the users that they can skip seat selection.

In order to start visualizing the interaction effort, we created an evaluation in UX-Analyzer for "Air somewhere" with two different versions, one for Deceptive and the other one for Normal. Each version's script was embedded in the corresponding page to send the interaction logs to UX-Analyzer. To gather user interaction data, a total of 152 participants (101 male, 51 female) completed a flight check-in on the website. They were recruited online via email and they were proportionally divided between the two versions. The task was online and self-moderated. Participants were given a reservation code and they had to fill in some personal information and decide between paying an extra for a seat or not.

Table 1 shows the resulting overall effort score and the specific score for the option "I don't want to select a seat" (Target Option Effort column). The table also contains the number of collected samples for each version. The total amount of the samples is 149 because there were 3 participants that could not complete the task.

Table 1: Interaction effort.

Version	#Users	**Overall Eff.**	Target Option Eff.
Deceptive	86	1.4	1.5
Normal	63	1.2	1.3

Comparing the results, the overall effort score of the Deceptive version is slightly greater than the one of the Normal version, which also happens when looking at the target option effort. Although this difference is not big enough to state that one version requires less effort than the other, there are interesting findings that could be useful for future improvements. For instance, the effort score of the "Continue" button (next to the target option) is 2 for Deceptive and 1 for Normal version. This may be explained by the fact that, in the Deceptive version, "Continue" is disabled if there is no seat selected, so there might have been users who tried to click on "Continue" without seeing the "I don't want to select a seat" option. Moreover, in the same seat selection page of Deceptive, the page's logo that takes the user to the website home has an effort of 2.5 (compared with 1.4 in the other version). A higher score in the Deceptive version may be an indicator of certain user confusion generated by not perceiving how to proceed without selecting a seat.

Another important aspect to mention is that the compared designs only differ in one specific element on the seat selection page, but the overall score for both versions also considers the widgets that are also present on the website's homepage (Figure 5) since the participants started the task from there. Although in this case we did not find any problematic widget on the homepage, there may be other cases in which users struggle with multiple widgets on different pages.

The above conjectures are examples of the analysis that can be stated besides just assessing which version results in a lower effort score. Definitely, it is necessary to conduct more evaluations to establish potential causes, but the tool offers a starting point as where to look for UX issues.

5 VALIDATION

In order to assess the applicability of UX-Analyzer in the context of a development process we conducted interviews with 5 UX experts that work in the industry. The interview consisted of 6 questions and it was divided in two parts. First we asked experts 2 questions targeted to determine the practices and tools that they use to evaluate web designs. Then, we presented a short introduction to the interaction effort score and a tool demo to gather their feedback with 4 specific questions. Finally, participants completed a two-items questionnaire (UMUX-Lite) to get an overall score of UX-Analyzer usability. This questionnaire has been proved to be highly correlated with the standard System Usability Scale (SUS). Next, we report the results of the interviews, focusing on the most relevant comments made by the participants.

5.1 Do You Use Analytics Tools on Your Web? Which Ones? When? If Not, Why Not?

This first question was intended to assess if participants use any kind of quantifiable metrics to get feedback from their designs. Out of the 5 participants, 3 answered that they use tools like Google Analytics[1] and Microsoft Clarity[2]. One of these experts commented that many times they have to put aside the analytics, to focus on new features that are required by the client. Regarding the 2 participants with a negative answer, one of them indicated that this is due to the lack of time to evaluate the metrics within the design process

[1]https://tagmanager.google.com/
[2]https://clarity.microsoft.com/

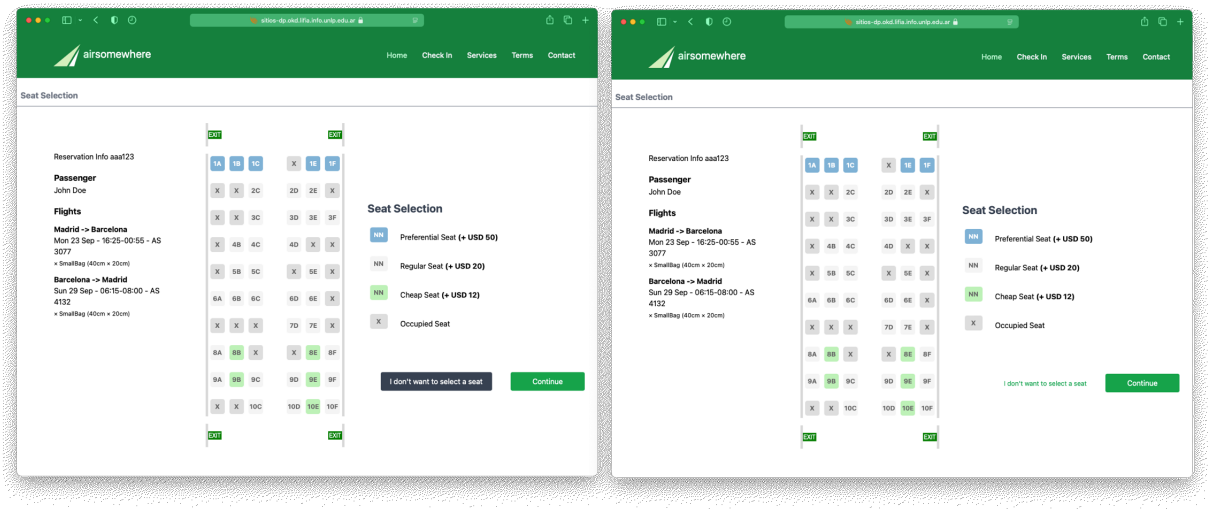

Figure 6: Test website variants, on the left the "Normal" option and on the right the "Deceptive" option.

that they follow, and the other one reported that analytics are not required by the client.

5.2 Short Intro of Interaction Effort Score

Before asking participants about the interaction effort, we explained them the main goal of the score and how it is calculated. We used an illustration with two alternative widgets and some micro-measures that are specific to them, so participants could easily understand the idea behind the score.

5.3 How Do You Usually Decide between Alternative Widgets or Designs? Do You Have the Chance to Use Objective Metric?

As the follow-up question suggests, this was intended to determine if UX designers use quantitative indicators that can be compared or related with the user interaction effort score. As a result, only one participant reported an objective metric which is the number of users clicks. Remaining participants indicated that these decisions are subjective, considering different aspects such as qualitative evaluations on prototypes, interviews with stakeholders, previous experience and design theory.

5.4 What Do You Think About the Interaction Effort Score?

All the participants agreed that having an objective score is ideal for facilitating decision-making. One participant stressed that the score can be specially useful to quickly gain feedback during the first phases of a product. The reason is that they have faced situations in which they do not have an argument to convince the client that they may be taking a wrong way. Another expert described cases in which they usually take into account design aesthetics and implementation effort to make a decision, so this metric can help to avoid loosing the user interaction perspective.

5.5 Would You Consider It When Making Design Decisions?

Answers to this question are somewhat overlapped with the previous one, since 100% of participants stated that they would take into account the score for decision making. Regarding the importance, participants indicated that the effort could be decisive when comparing alternative designs, specially in those teams that do not have enough resources to conduct other types of evaluations.

5.6 UX-Analyzer Demonstration

In this step we presented a concrete example of the interaction effort using UX-Analyzer. Particularly, we showed the evaluation previously described in section 4 with the two alternative versions. Web pages illustrated in Figure 6 were also shown to give participants context about what we were evaluating. The tool demo focused on the different perspectives of the effort score (overall, user sessions and widgets) and the conjectures made in the use case.

5.7 Would You Adopt the Tool?

This question aims at assessing the overall usability of UX-Analyzer besides the interaction effort score itself. All the participants agreed that they would use UX-Analyzer, although 3 of them mentioned some design aspects that should need to be improved. The most important one is to incorporate visual elements to the reports such as graphics and dashboards, so it is easier for them to understand the different scores.

5.8 What Changes Would You Make to UX-Analyzer?

The last question was intended to give participants the opportunity to make comments regarding new features that could be added to the tool or any other opinion that they consider relevant. Besides designs improvements mentioned in the previous question, the answers included the following features or nice-to-have:

- Allow users to set target scores for specific versions under test and warn them when the real score is out of indicated threshold or values.
- Add a description in the tool about the possible values that the effort score can take, so the users can have a better grasp if the results are good or bad.
- Enable downloading reports to share them with the client or other interested stakeholders. This feature is in line with presenting the results in a more visual friendly way.
- The way that the widgets weight is presented makes difficult to estimate the importance that is assigned to each element. It would be appropriate to find a way to standardize the weights to allow replicating the same value on different widgets.

5.9 UMUX-Lite Questionnaire

Questionnaire scores were transformed into a 0 to 100 scale for a better interpretation. Results showed an average score of 55/100 which is an indicator of good usability, but at the same time it shows that there is room for improvements as suggested by participants.

6 DISCUSSION

In the previous section we described a use case of the tool to visualize the user interaction effort score of alternative versions for a specific websites, and we gave examples of the type of conclusions that we expect to draw from UX-Analyzer.

The analysis performed in the use case may be limited due to the fact that effort scores obtained for the alternative versions did not show a clear difference, being both close to 1. We believe that this is related to the task design, which was simplified to be done in a self-moderated way, and it was completed out of its real context. A concrete example of this context is that the user has to buy a ticket before being able to check-in for a flight. Probably, performing the task in its real context adds more complexity and this is reflected in a greater effort score, which in turn could give more insights about the problems that users may experience using the website. Moreover, since the interaction effort is a metric based on individual UI widgets, it can provide better results when users freely navigate through the target application without being constrained to a specific task or user test.

Concerning the reports provided by UX-Analyzer, besides comparing alternative versions it gives the possibility to observe the interaction effort aggregated by each interactive widget under analysis, which may be present on different pages. This information is helpful to identify the individual elements that can cause problems to the users and it can be a decisive aspect when comparing alternatives as suggested by interviewed UX designers. Disabling the widgets and adjusting their importance is also a relevant feature that allows the user to focus the analysis on specific elements. The widgets weight concept is something that needs further improvements to show what its influence is in the overall effort score, and apply the same widget settings to different versions.

Another report available in the tool is the average interaction effort that results from each user session logged. During the previous use case we could not extract any relevant information about the interaction effort at the user session level. When UX designers analyzed this report, they stated that with the effort score and session time is difficult to assess what happened with a particular user. In this regard, we plan to include more details about each user session which includes a screen recording of it, so the user of UX-Analyzer can observe specific cases in which a user of the analyzed website is struggling with the user interface.

The validation with UX designers allowed us to draw relevant conclusions about UX-Analyzer's applicability and potential improvements. One of the main outcomes from the interviews is that UX designers sometimes do not use analytics, and even when they use them, they do not have enough time to analyze the results and take action. They mainly rely on qualitative assessments and previous experience when they have to decide between alternative designs. In this context, designers found the interaction effort very useful not only for making decisions but also for communication with stakeholders.

Another valuable outcome from the interviews is that all the participants would adopt UX-Analyzer. They also gave us feedback on how to boost the tool's user experience as well as new features that can facilitate the analysis.

7 CONCLUSIONS AND FUTURE WORK

In this work we showed UX-Analyzer, a web tool to evaluate web pages with respect to the user interaction effort. The tool provides different perspectives of the user interaction effort, giving the chance to visualize a single effort score for each analyzed version, as well as the effort of each user session and the average effort demanded by each specific widget.

To evaluate the tool under real contexts of use, we conducted 2 studies. On one hand, a case study with real users in which we analyzed the resulting effort scores. This analysis pointed us to possible improvements for the websites. On the other hand, we conducted a survey with UX professionals, which allowed us to assess their current practices regarding UX metrics, their perception of the tool and the potential for its adoption.

We are currently working on improving UX-Analyzer. Most importantly, we intend UX-Analyzer to show other metrics besides interaction effort, so we can evaluate other aspects of UX. In this way, the tool could be used as a dashboard to inspect and monitor different metrics that help UX experts or other interested team member to keep the UX under control. This will also allow us to study the potential relationship between interaction effort and other ways of measuring UX.

ACKNOWLEDGMENTS

The authors wish to acknowledge the support from the Argentinian National Agency for Scientific and Technical Promotion (ANPCyT), grant number PICT-2019-02485.

REFERENCES

[1] Manal M Alhammad and Ana M Moreno. 2022. Integrating user experience into Agile: an experience report on lean UX and Scrum. In *Proceedings of the ACM/IEEE 44th International Conference on Software Engineering: Software Engineering Education and Training*. Association for Computing Machinery, New York, NY, USA, 146–157.

[2] Omar Badran and Shafig Al-Haddad. 2018. The impact of software user experience on customer satisfaction. *Journal of Management Information and Decision Sciences* 21, 1 (2018), 1–20.

[3] Maxim Bakaev, Tamara Mamysheva, and Martin Gaedke. 2017. Current trends in automating usability evaluation of websites: Can you manage what you can't measure?. In *Proceedings - 2016 11th International Forum on Strategic Technology, IFOST 2016*. IEEE, New York, NY, USA, 510–514. https://doi.org/10.1109/IFOST.2016.7884307

[4] Lara Bertram and Markus Dahm. 2022. Conceptual design and implementation of an automated metrics and model-based usability evaluation of UI prototypes in Figma. In *Mensch und Computer 2022 - Workshopband*. Gesellschaft für Informatik eV, 53175 Bonn, Germany, 1–6.

[5] Tanja Blascheck, Kuno Kurzhals, Michael Raschke, Michael Burch, Daniel Weiskopf, and Thomas Ertl. 2017. Visualization of eye tracking data: A taxonomy and survey. *Computer Graphics Forum* 36, 8 (2017), 260–284.

[6] Sara Bouzit, Calvary Gaelle, and Jean Vanderdonckt. 2016. Automated Evaluation of Menu by Guidelines Review. In *RoCHI - International Conference on Human Computer Interaction*. MATRIX ROM, Bucharest, Romania, 1–12.

[7] Raluca Budiu. 2013. Interaction Cost. https://www.nngroup.com/articles/interaction-cost-definition/. Date accessed: 12/28/2023.

[8] Hendrik Bünder and Herbert Kuchen. 2019. Towards behavior-driven graphical user interface testing. *ACM SIGAPP Applied Computing Review* 19, 2 (2019), 5–17.

[9] Stuart K Card, Thomas P Moran, and Allen Newell. 1980. The keystroke-level model for user performance time with interactive systems. *Commun. ACM* 23, 7 (1980), 396–410.

[10] Fang Chen, Natalie Ruiz, Eric Choi, Julien Epps, M. Asif Khawaja, Ronnie Taib, Bo Yin, and Yang Wang. 2012. Multimodal behavior and interaction as indicators of cognitive load. *ACM Transactions on Interactive Intelligent Systems* 2, 4 (2012), 1–36. https://doi.org/10.1145/2395123.2395127

[11] Fang Chen, Jianlong Zhou, Yang Wang, Kun Yu, Syed Z. Arshad, Ahmad Khawaji, and Dan Conway. 2016. Theoretical Aspects of Multimodal Cognitive Load Measures. In *Robust Multimodal Cognitive Load Measurement*. Springer International Publishing, Cham, 33–71. https://doi.org/10.1007/978-3-319-31700-7_3

[12] Tiago Silva Da Silva, Milene Selbach Silveira, Frank Maurer, and Fábio Fagundes Silveira. 2018. The evolution of agile UXD. *Information and Software Technology* 102 (2018), 1–5.

[13] Qi Dou, Xianjun Sam Zheng, Tongfang Sun, and Pheng-Ann Heng. 2019. Webthetics: Quantifying webpage aesthetics with deep learning. *International Journal of Human-Computer Studies* 124 (2019), 56–66. https://doi.org/10.1016/j.ijhcs.2018.11.006

[14] Peitong Duan, Casimir Wierzynski, and Lama Nachman. 2020. Optimizing User Interface Layouts via Gradient Descent. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376589

[15] DugWood. 2011. ClickHeat, the tool that warms your visitors' mouse activities. https://www.dugwood.com/clickheat/index.html Accessed = 2024-11-25.

[16] Sergio Firmenich, Alejandra Garrido, Julián Grigera, José Matías Rivero, and Gustavo Rossi. 2019. Usability improvement through A/B testing and refactoring. *Software Quality Journal* 27 (2019), 203–240.

[17] Paul M. Fitts. 1954. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology* 47, 6 (1954), 381–391.

[18] Juan Cruz Gardey, Julian Grigera, Andres Rodriguez, and Alejandra Garrido. 2024. UX-Analyzer: Visualizing Interaction Effort for Web Analytics. In *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing (SAC '24)*. Association for Computing Machinery, New York, NY, USA, 1774–1780. https://doi.org/10.1145/3605098.3636013

[19] Juan Cruz Gardey., Julián Grigera., Andrés Rodríguez., Gustavo Rossi., and Alejandra Garrido. 2022. An Interaction Effort Score for Web Pages. In *Proceedings of the 18th International Conference on Web Information Systems and Technologies - WEBIST,*. INSTICC, SciTePress, Setubal, Portugal, 439–443. https://doi.org/10.5220/0011591400003318

[20] Juan Cruz Gardey, Julián Grigera, Andrés Rodríguez, Gustavo Rossi, and Alejandra Garrido. 2022. Predicting interaction effort in web interface widgets. *International Journal of Human-Computer Studies* 168 (2022), 102919. https://doi.org/10.1016/j.ijhcs.2022.102919

[21] Julián Grigera, Juan Cruz Gardey, Andres Rodriguez, Alejandra Garrido, and Gustavo Rossi. 2019. One metric for all: Calculating interaction effort of individual widgets. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–6.

[22] Julián Grigera, Alejandra Garrido, José Matías Rivero, and Gustavo Rossi. 2017. Automatic detection of usability smells in web applications. *International Journal of Human-Computer Studies* 97 (2017), 129–148.

[23] Marc Hassenzahl and Noam Tractinsky. 2006. User experience-a research agenda. *Behaviour & information technology* 25, 2 (2006), 91–97.

[24] Andreas Hinderks, Dominique Winter, Martin Schrepp, and Jörg Thomaschewski. 2019. Applicability of user experience and usability questionnaires. *Journal of Universal Computer Science, 25 (13), 1717-1735.* 25, 13 (2019), 1717–1735.

[25] ISO. 2019. *ISO 9241-210:2019 - Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems*. ISO/TC 159/SC 4. https://www.iso.org/standard/77520.html

[26] Ron Kohavi and Roger Longbotham. 2017. Online Controlled Experiments and A/B Testing. *Encyclopedia of machine learning and data mining* 7, 8 (2017), 922–929.

[27] Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and evaluation of a user experience questionnaire. In *Symposium of the Austrian HCI and usability engineering group*. Springer, Springer Berlin Heidelberg, Berlin, Heidelberg, 63–76.

[28] James R Lewis, Brian S Utesch, and Deborah E Maher. 2013. UMUX-LITE: when there's no time for the SUS. In *Proceedings of the SIGCHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 2099–2102.

[29] Jiajia Li, Grace Ngai, Hong Va Leong, and Stephen C. F. Chan. 2016. Multimodal human attention detection for reading from facial expression, eye gaze, and mouse dynamics. *SIGAPP Appl. Comput. Rev.* 16, 3 (2016), 37–49. https://doi.org/10.1145/3015297.3015301

[30] Yang Li, Samy Bengio, and Gilles Bailly. 2018. Predicting human performance in vertical menu selection using deep learning. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3173574.3173603

[31] Laura Luther, Victor Tiberius, and Alexander Brem. 2020. User Experience (UX) in business, management, and psychology: A bibliometric mapping of the current state of research. *Multimodal Technologies and Interaction* 4, 2 (2020), 18.

[32] Jevgeni Marenkov, Tarmo Robal, and Ahto Kalja. 2018. Guideliner: A tool to improve web UI development for better usability. In *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*. Association for Computing Machinery, New York, NY, USA, 1–9.

[33] Diogo S. Martins, Bruna C. R. Cunha, Cristiane A. Yaguinuma, Isabela Zaine, and Maria da Graça C. Pimentel. 2019. Effects of interactive video annotations on students' browsing behavior and perceived workload. *SIGAPP Appl. Comput. Rev.* 19, 2 (Aug. 2019), 44–57. https://doi.org/10.1145/3357385.3357389

[34] Anna-Lena Meiners, Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. 2023. A Benchmark for the UEQ+ Framework: Construction of a Simple Tool to Quickly Interpret UEQ+ KPIs. *International Journal of Interactive Multimedia and Artificial Intelligence* 9, 1 (2023), 104–111.

[35] Eleni Michailidou, Sukru Eraslan, Yeliz Yesilada, and Simon Harper. 2021. Automated prediction of visual complexity of web pages: Tools and evaluations. *International Journal of Human-Computer Studies* 145 (2021), 102523. https://doi.org/10.1016/j.ijhcs.2020.102523

[36] Abdallah Namoun, Ahmed Alrehaili, and Ali Tufail. 2021. A review of automated website usability evaluation tools: Research issues and challenges. In *International Conference on Human-Computer Interaction*. Springer International Publishing, Cham, 292–311.

[37] Divya Natesan, Morgan Walker, and Shannon Clark. 2016. Cognitive bias in usability testing. In *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care*, Vol. 5. SAGE Publications, Los Angeles, CA, 86–88.

[38] Jakob Nielsen. 2005. Putting A/B Testing in Its Place. https://www.nngroup.com/articles/putting-ab-testing-in-its-place/.

[39] Harkirat Kaur Padda. 2003. *QUIM map: a repository for usability/quality in use measurement*. Ph.D. Dissertation. Concordia University.

[40] Fabio Paternò, Antonio Giovanni Schiavone, and Antonio Conti. 2017. *Customizable Automatic Detection of Bad Usability Smells in Mobile Accessed Web Applications*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3098279.3098558

[41] Virpi Roto, EL-C Law, Arnold POS Vermeeren, and Jettie Hoonhout. 2011. User experience white paper: Bringing clarity to the concept of user experience. In *Outcome of Dagstuhl Seminar 10373: Demarcating User Experience*. s.n, s.n, 1–12.

[42] Jeff Sauro. 2015. SUPR-Q: A comprehensive measure of the quality of the website user experience. *Journal of usability studies* 10, 2 (2015), 68–86.

[43] Maximilian Speicher, Andreas Both, and Martin Gaedke. 2014. Ensuring Web Interface Quality through Usability-Based Split Testing. In *Web Engineering*. Springer International Publishing, Cham, 93–110. https://doi.org/10.1007/978-3-319-08245-5_6

[44] Desirée Sy. 2007. Adapting usability investigations for agile user-centered design. *Journal of usability Studies* 2, 3 (2007), 112–132.

[45] Hongyan Wan, Wanting Ji, Guoqing Wu, Xiaoyun Jia, Xue Zhan, Mengting Yuan, and Ruili Wang. 2021. A novel webpage layout aesthetic evaluation model for quantifying webpage layout design. *Information Sciences* 576 (10 2021), 589–608. https://doi.org/10.1016/j.ins.2021.06.071

[46] Norhanisha Yusof, Nor Laily Hashim, and Azham Hussain. 2022. A Conceptual User Experience Evaluation Model on Online Systems. *International Journal of Advanced Computer Science and Applications* 13, 1 (2022), 1–11.