SADIO Electronic Journal of Informatics and Operations Research (EJS) Vol. 25 No. 1 (2026) e-ISSN 1514-6774
https://revistas.unlp.edu.ar/ejs | https://doi.org/10.24215/15146774e092

How Mobile UX Smells Affect Interaction Efficiency: A Multi-Metric Empirical Study

Claudio Raverta[1], Julián Grigera[1,2,3][0000-0002-7962-4312], Juan Cruz Gardey[1][0000-0002-1765-8189], and Alejandra Garrido[1,2][0000-0002-5052-705X]

[1] LIFIA, Fac. de Informatica, Univ. Nac. de La Plata, La Plata, Argentina
{craverta,juliang,jcgardey,garrido}@lifia.info.unlp.edu.ar
[2] CONICET, Argentina
[3] CICPBA, Pcia. de Buenos Aires, Argentina

Abstract. The concept of UX smells has been recently studied as a systematic way to detect predefined user interaction issues and fix them with cataloged solutions. Most of the existing literature about UX smells focuses on desktop web applications, while there are only a few works addressing the mobile web. Although specific UX smells for mobile interactions have been proposed, there are no objective evaluations to determine their impact on the perceived UX. In this work, we evaluated 6 mobile UX smells (3 from the literature and 3 new proposals) with respect to efficiency in use. We conducted an online evaluation with 72 participants in 3 real websites, each one with a set of specific mobile UX Smells. In this evaluation, we compared each website to a refactored version of itself, i.e. with proposed fixes for each of the smells. To do this, we ran a between-subject experiment in which participants completed 10 everyday tasks on the websites while we measured their efficiency in terms of task completion time and number of user interaction events. As a complementary post-hoc analysis, we also grouped temporally close interaction events into interaction bursts, providing an additional efficiency-related perspective. All the captured metrics were compared in the default version of the websites vs. their refactored counterparts. Results showed that in most cases (15/20), either the time to complete the task or the amount of interaction events were higher in the presence of UX smells. Moreover, in 7 of the cases, the observed differences were statistically significant ($p<0.05$). The burst-based analysis was consistent with these trends.

Keywords: UX Smells, Mobile Web, Efficiency in Use

1 Introduction

Despite the attention that User Experience (UX) has received in recent years, it continues to be a concern in Web User Interfaces (Web UIs) (Weichbroth, 2020). In particular, the mobile web suffers from problems related to its specific interaction style that differs from desktop UIs - e.g. smaller displays, no mouse cursor, shorter attention span. These mobile UX issues have been studied and

Received August 2025; Accepted November 2025; Published February 2026

cataloged under different taxonomies and names in order to help practitioners and researchers identify them and also solve them (Limaylla Lunarejo et al., 2020; Paternò et al., 2016) . Some of these studies have proposed describing these issues as Usability Smells or the more general UX Smells (Gardey et al., 2020). The concept of “smell" is a well-known metaphor in software engineering (taken from the original Code Smell denomination), which hints an issue solvable by some cataloged solutions called refactorings (Buono et al., 2020; Gardey et al., 2020; Harms & Grabowski, 2014). Some examples of UX Smells that are specific to the mobile web are *Too Small or Close Element* (Nebeling et al., 2013), *Bad Readability* (Paternò et al., 2017) and *Page Expands Horizontally* (Limaylla Lunarejo et al., 2020).

Existing evaluations in the area have been conducted (de Santana & Baranauskas, 2015; Harms et al., 2014; Paternò et al., 2017), which add to the corpus of evidence, but they rather focus on the detection of smells by analyzing interaction patterns, and not on the impact that these smells may have. This could provide relevant information for web UI designers.

Regarding the evaluation of UX Smells impact, the literature proposes different approaches, but in general, evaluate the three classical aspects of usability, namely effectiveness, efficiency and satisfaction. However, these aspects are measured in laboratory settings, where satisfaction and effectiveness are often compromised by different kinds of bias (e.g. courtesy bias or completeness bias, respectively). In this work, we focus on efficiency. Capturing this aspect is less intrusive, which means, on one hand, that it tends to be more ecologically valid than effectiveness, and on the other hand more objective than satisfaction. Additionally, even if it is only a partial metric for usability, efficiency was found to be the most frequently identified aspect in usability definitions in the literature, according to a 2020 SLR (Weichbroth, 2020).

When measuring efficiency, most evaluations focus on completion times, which is a widely accepted metric for this aspect, and at the same time also easy to capture. However, time alone could be misleading since, for instance, interaction is often interrupted for external factors. Also, users may have different interaction speeds leading to different times measurements even if the relative efficiency is similar for them. For these reasons, we complement time measurements with event logs, in order to capture other aspects of efficiency. These events were designed to capture standard actions like screen touches or scolling actions, but also mobile-specific gestures like pinch-zoom.

In this work, we describe an evaluation conducted with 70+ participants to answer our primary research question: **how does the presence of mobile UX smells affect efficiency?** As previously mentioned, we measured efficiency using completion times and event counts. Additionally, we complemented the event count study with an additional post-hoc analysis in which we counted events grouped in units or "bursts" of interaction. Our evaluation was conducted on real websites with real UX Smells, while in the baseline ones we corrected these smells using UX refactoring. We designed this setting to improve the external validity of our results.

We used three different snapshots of real websites, and evaluated the following UX smells: *Too Small or Close Element*, *Bad Readability*, *Lost Search*, *Misplaced Menu*, *Unclear Form and Unclear Widget*.

Results indicate that, in most cases, mobile websites suffering from these UX smells take users a statistically significant longer time, or higher number of interaction events. In the remaining cases, even if no statistically significant differences were found, the averages are still higher. In the case of interaction events, the complementary study of events bursts revealed that the count of fast-paced sequences of action considered as units, are also generally higher on sites that contain UX smells. The contributions of this work are twofold: on the one hand, the experiment results represent new evidence on the impact of UX Smells in mobile devices. On the other hand, 3 new mobile UX Smells are defined, which have also been tested in the experiment.

The remaining sections of this paper examine: related work, design and implementation including details about participants, UX smells and platform, and results based on the analysis of data collected in the experimental phase.

2 Related work

The UX smell concept was introduced many years ago to describe UI designs that can provide bad user experience. UX smells are inspired in Code smells, the difference is that the former are focused on the external system quality, while the latter analyze internal code standards. The first works about UX smells reported several hints of poor interface design, specially in web applications (Almeida et al., 2015; Garrido et al., 2013; Harms & Grabowski, 2014). Harms et al. proposed a method to identify a set of 5 usability smells by recording the users actions and analyzing the generated task tree models (Harms & Grabowski, 2014). In one of our previous works, we presented a catalog of usability smells that can be detected automatically through the analysis of user interaction events (Grigera et al., 2017). Each of these smells may be solved by applying usability refactorings, which are predefined transformations that improve the user interaction without changing the underlying functionality. Having cataloged smells is not only useful to easily find potential issues, but also facilitates their resolution. The usability smell concept was then re-framed to UX smell, motivated by the need of including other UX aspects beyond usability such as aesthetics and preferences (Buono et al., 2020; Gardey et al., 2020). Therefore, throughout this paper we use the term UX smell to avoid limiting the analysis to usability.

The widespread use of mobile devices has shifted the attention from desktop to mobile user interfaces. Although they still support the web, mobile devices have changed the way of interacting with applications. In this regard, some investigations analyzing UX smells for mobile user interfaces have been conducted (Limaylla Lunarejo et al., 2020; Paternò et al., 2017). Paternò et al. proposed six UX smells related with limited screen size and touch inputs (Paternò et al., 2017). They also developed a tool to automatically detect them by analyzing the interaction logs captured on the device's browser. In the same way, Limaylla et al. presented another method based on user interaction analysis to detect 2 proposed UX smells together with 4 smells previously presented by Paternò et al. (Paternò et al., 2017).

A more recent work by Ali et al. (Ali et al., 2024) presents M-UI-R, an approach for identifying mobile UI smells in mobile Apps. M-UI-R uses Convolutional Neural Networks (CNNs) trained on history and real-time data for the detection, achieving over 70% in both precision & recall. Another similar approach is UIS-Hunter (Yang et al., 2021), which uses computer-vision techniques to detect UI guideline violations. These approaches are relevant to our work, but there are some differences: one is that they focus on Apps instead of the mobile web, but most importantly, their definition of UI smell is different from the one we use. What these authors regard as a UI smell is related to visual issues like wrong layout or overlapping widgets, which in some cases could even be considered as bugs. Additionally, these issues are strictly visual and do not consider user interaction.

Most of the works mentioned before were aimed at providing new UX smells and detecting them automatically. For that reason, the validations conduced in these researches are mainly concerned with assessing the effectiveness of the tools to identify the proposed smells. Although some of them include qualitative assessments of the UX smells utility, they lack objective evaluations to analyze the impact of the UX smells on the user experience. As far as we know, our research is the first one evaluating how the presence of UX smells influence on quantitative UX metrics.

3 Experiment

We designed an online evaluation to study the behavior of users when facing daily tasks in the mobile web. The primary research question is "how does the presence of mobile UX smells affect efficiency?". Thus, we posed 2 hypotheses to assess task efficiency in different conditions:

- H1: Users completing simple tasks on a mobile website with known UX smells **take more time**.
- H2: Users completing simple tasks on a mobile website with known UX smells **produce more interaction events**.

To test these hypotheses, we captured the following 2 dependent variables: completion time and interaction events count. For the capture, we developed a tool that also guided the participants throughout the tasks in each of the 2 versions of the 3 test websites. The independent variable was the version of the website provided to the participants, i.e., with or without UX Smells. Since most UX smells were present on each of the 3 websites, we were able to assess their influence in more than one setting.

3.1 Materials

The UX smells we tested were originally found in the websites chosen for the tests, from which we prepared clones with pruned navigation paths. Moreover, we created clean versions of the test sites by refactoring the UX smells. Out of the 6 UX Smells tested, 3 of them were previously defined in the literature and the remaining ones were defined in this work.

The following is a list of the tested UX smells:

- **Too Small or Close Element**: cataloged in earlier research (Nebeling et al., 2013; Paternò et al., 2017). Indicates that a clickable element has a very small clickable area, typical in "close ad" buttons.
- **Bad Readability**: cataloged by Paternò et al. (Paternò et al., 2017). As its name suggests, it signals small fonts or generally difficult to read text. In our experiment, we specifically test squashed columns, which affect readability by having very little text (as little as one word) per line.
- **Distant Content**: cataloged by Paternò et al (Paternò et al., 2017) and Grigera et al. (Grigera et al., 2017). It Describes elements whose display or interaction is crucial for the funtionality of the app but are placed in such a way that the user is forced to make repeated scrolls up or down to reach them. Among the chosen websites it consists mainly of search boxes or forms.
- **Misplaced Menu**: Describes a menu that does not follow the standardized location or aspect for mobile web menus, i.e. "burger" icon, and top left or top right placement. It is derived from some of the usability guidelines listed by Alonso-Virgós et al. (Alonso-Virgós et al., 2020) which highlight the importance of uniformity and consistency in the design, such as respecting conventional layouts and proper use of common icons.
- **Unclear Form**: Describes forms whose design does not clearly highlight their different components or their location in relation to the rest of the page elements. This was studied as a side effect of Flat Design (Burmistrov et al., 2015).
- **Unclear Widget**: Describes functional elements such as links and buttons whose use of typography and colors prevent their distinction from regular elements of the site.ide The influence of these properties has been studied as an issue of smartphone-specific affordance (Burmistrov et al., 2015; Chen et al., 2024). Also, size relationship with the rest of the site may be designed for desktop versions and does not contemplate the specific needs of mobile navigation.

The three websites used were offline clones of publicly accessible domains: health insurance, news and retail. IOMA is the official website of a public medical insurance institute from Buenos Aires, the most populated province in Argentina. It mainly displays information regarding medical attention options for its affiliates and public healthcare campaigns. Balcon Plantas (currently offline) was an online shop specialized in plants, plant care tools and accessories. Kabytes is a blog dedicated to graphic design, web programming and social media. The test websites can be seen in Figure 1 and accessed in a public repository[4].

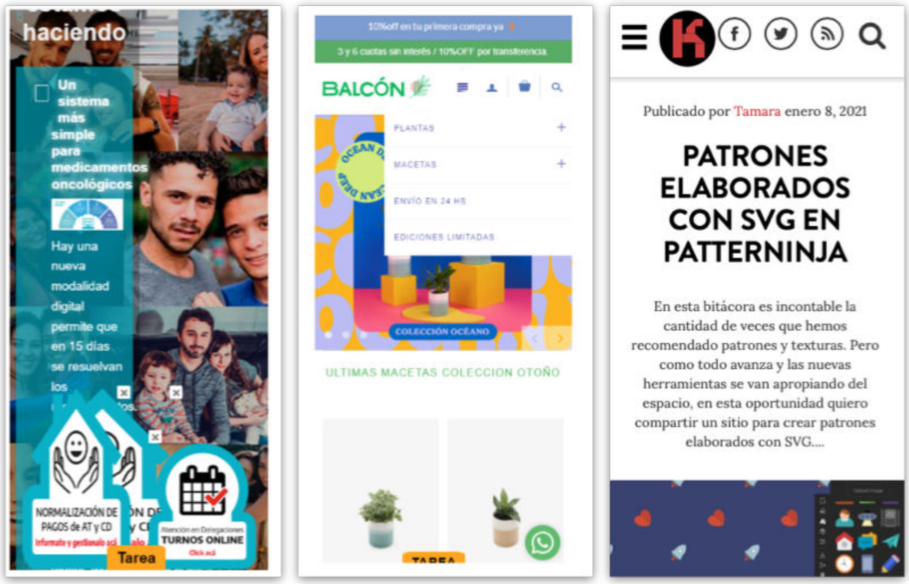

Fig. 1. Screenshots of the test websites. From left to right: IOMA, Balcón Plantas, Kabytes. All accessed March 2024.

Early on in the process, we decided to select publicly available websites with real UX problems, so that the websites could be left intact with no additional manipulation from the authors, thus ensuring fidelity to real-life environments. Therefore, the selection is based on their popularity, and also for the presence of UX smells reported in the literature, representative of common problems in mobile browsing. All sites were selected throughout 2023 and saved into offline versions to be included in the experiment.

The resulting smells range from patently obvious design blunders with an expected effect on UX, to subtle imperfections that deviate from good practices but have an uncertain influence on the overall experience.

For each site we implemented two versions: a first version "as is", including the original UX smells and a corrected versions, which went through an editing process.

The tasks were designed to replicate typical situations faced by a user when browsing from a mobile device. They covered a wide range of interactions with interface elements, such as: accessing an interactive menu and selecting an item, searching and selecting an item in the text, registering via email, using a search

[4] https://github.com/claudiorave/mobile_usability_scripts

bar, etc. Table 1 describes the tasks for each of the sites (translated from original Spanish).

Table 1. Description of tasks provided to participants (translated from original Spanish) and related UX Smells.

Task Nº	Website	Task Description	UX Smell
1	IOMA	Close 3 pop-up banners that appear at the bottom of the screen	*Too Small Element*
2	IOMA	Find a keyword within a text in a columnar layout that makes it difficult to read	*Bad Readability*
3	IOMA	Find the search box and search for a keyword	*Distant Content*
4	Balcón Plantas	Find the main menu and select one of its items	*Misplaced Menu*
5	Balcón Plantas	Find the search box and search for a keyword	*Distant Content*
6	Balcón Plantas	Find the subscription form and enter a valid email address	*Unclear Form*
7	Balcón Plantas	Find key text on the page linked to an image	*Unclear Widget*
8	Kabytes	Find the main menu and select one of its items	*Misplaced Menu*
9	Kabytes	Below each news review there is a 'Read more' button. Activate any 3 out of a total of 10.	*Unclear Widget*
10	Kabytes	Find the search box and search for a keyword	*Distant Content*

3.2 Platform and metrics

The platform consists of two complementary web applications. One is used to host the test circuit provided to participants and the other stores data and displays it for analysis with the aid of visual tools.

The event types that were captured comprise: **clicks**, **misclicks**, **scrolls**, **orientation change** and **pinch zoom**. **Clicks** and **misclicks** refer to mobile devices' screen touch event, the first one consisting in those involved directly in the resolution of the task in progress and the latter for the remainder of the touches. These are particularly useful to detect potential touches that missed the target. **Scrolling** describes the action of dragging the page vertically or horizontally so that the screen focuses on another portion of the site, in the same way that a mouse wheel or a lateral scroll bar operate in a desktop browser. Finally, **pinch zoom** and **orientation change** are two events with no direct equivalent in desktop navigation and both are linked to the need to focus on a portion of the site to improve its visualization. In the first one, "pinching" on the required area to zoom in on it, while in the latter, rotating the device to obtain a landscape view of the section.

A script implemented in the client side handles the detection of events and sends them into the logging application together with relevant data like its timestamp and associated HTML elements. While the present work focuses only on event counts, we collected this contextual data for its potential uses in further research.

The test circuit was developed entirely in HTML, JavaScript and CSS, while Python/Django, Django REST Framework, Django Channels, PostgreSQL were used for the logging system. The procedure first displays a screen with instructions and the user is then guided by interactive visual elements that appear as each task is completed, and can be consulted at any time by pressing a help button, visible at all times on the screen. Figure 2 shows screen captures of the circuit tool as seen by the participants.

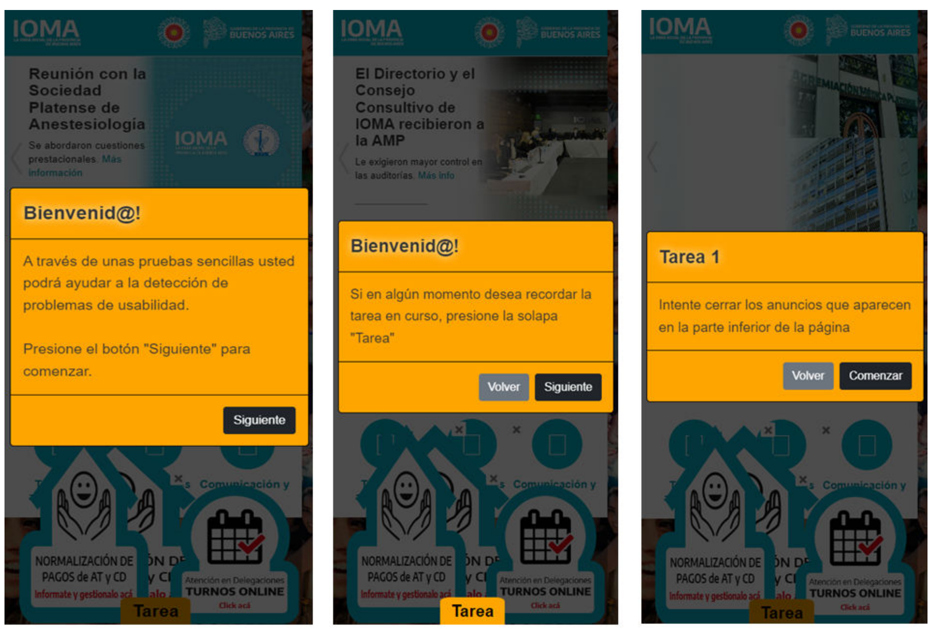

Fig. 2. Welcome modal with instructions, task remembering function and first task.

Each event performed by the user in the test circuit is sent via HTTP requests to the logging system and stored in a relational database under the same session, an entity with a unique alphanumeric token that groups the events of each user within its circuit. In order to analyze the results, the logging system has a visual interface capable of representing the obtained data under different criteria, both at an individual level to examine in detail the sequences of events occurring in each session, and at a general level comprising all the user experiments to identify generalizable trends.

The tools used to capture events are available in a separate repository[5].

3.3 Participants and Execution

Participants were recruited online, since the experiment was self moderated. The call was spaced over a period of months throughout 2023, with new participants being invited as they became available. The invitation consisted of sending a link to the testing platform along with a brief description of the nature of the project to groups of contacts via instant messaging services, leaving aside instructions relating to the experiment that would be provided to the user once access to the testing site had been gained. We recruited a total 72 participants who completed the test. All participants were asked to use a mobile device (smartphone, not tablet), and their ages ranged between 20 and 55.

Since this was a between-subject design, each participant was assigned 1 of 2 possible versions of the task circuit. Circuit "A" consisted in going through the 3 sites in the following order: IOMA (refactored), Balcón Plantas (with smells), Kabytes (refactored). Circuit "B" reversed the versions, as follows: IOMA (with smells), Balcón Plantas (refactored), Kabytes (with smells). This way, each participant only using each website only once to counter a potential learning effect that would have been too significant a threat - hence the between-subject design. As soon as the participants entered the experiment, an algorithm implemented at the start of the application determined with equal probability the target version for each user. The user was then guided through the tasks on the test sites, and at all times they could see visual cues indicating task completion, which could be seen by pressing a help button visible at all times on the screen.

4 Results

In this section we present the results of the statistical tests for both studied variables, and also a post-hoc analysis of the raw data. We also describe the potential threats to validity.

4.1 Statistical Tests

We ran statistical tests to support our hypotheses. Since normality in the data could not be verified, we chose Mann Whitney, a non-parametric test. The results for the tests for each task are shown in Table 2.

The null hypotheses were rejected in 8/20 cases: 4 instances of H1 (time) and 4 instances of H2 (events count). However, in one case (Task 5) the rejected H0 for completion time **favored the variant with UX Smells**. That is, the participants completing task 5 took significantly more time with the refactored

[5] https://github.com/claudiorave/mobile_usability

Table 2. Results of Mann Whitney tests for each task. Rejected H0 in bold.

Task	Site	UX Smell	p-value time	p-value events
1	IOMA	Too Small or Close Element	**0.000097**	**0.000004**
2	IOMA	Bad Readability	0.276700	**0.003624**
3	IOMA	Lost Search	0.093860	0.093860
4	Balcón plantas	Misplaced Menu	**0.041750**	0.766500
5	Balcón plantas	Lost Search	0.229600	***0.000035***
6	Balcón plantas	Unclear Form	0.603900	0.525300
7	Balcón plantas	Unclear Widget	0.072000	**0.001427**
8	Kabytes	Misplaced Menu	**0.031210**	0.058820
9	Kabytes	Unclear Widget	0.869200	0.426000
10	Kabytes	Lost Search	**0.012620**	0.906900

version. Only for task 1, "Too Small or Close Element" smell was the null hypotheses rejected in both variables. Another result worth noticing were tasks 3 (Lost search - IOMA) and 9 (Unclear Widget - Kabytes) that did not reject any hypothesis.

4.2 Post-hoc Analysis

To dig deeper into the obtained results, we also analyzed the raw results for the measures of central tendency. Figures 3 (time variable) and 4 (events count variable) show boxplots that describe the data.

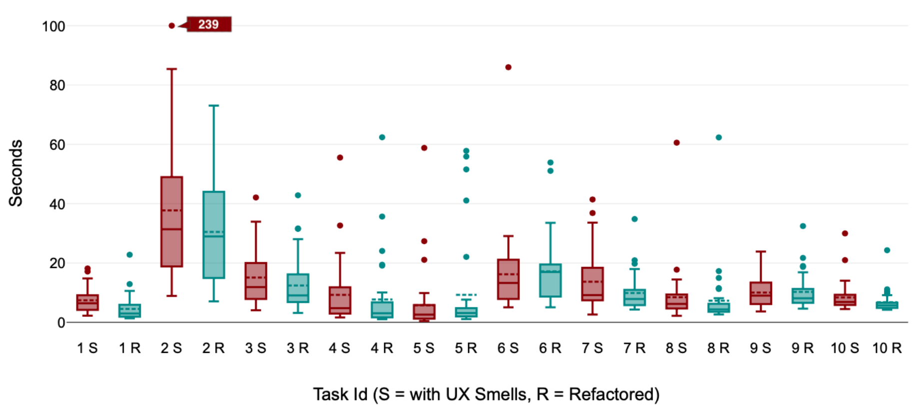

Fig. 3. Refactored and original (with smells) versions - Completion time per task.

As the charts show, besides the cases in which the null hypotheses were rejected, most average values are still higher in the original versions with UX Smells (15/20). There are however 5 exceptions, most notably task 5 (Search

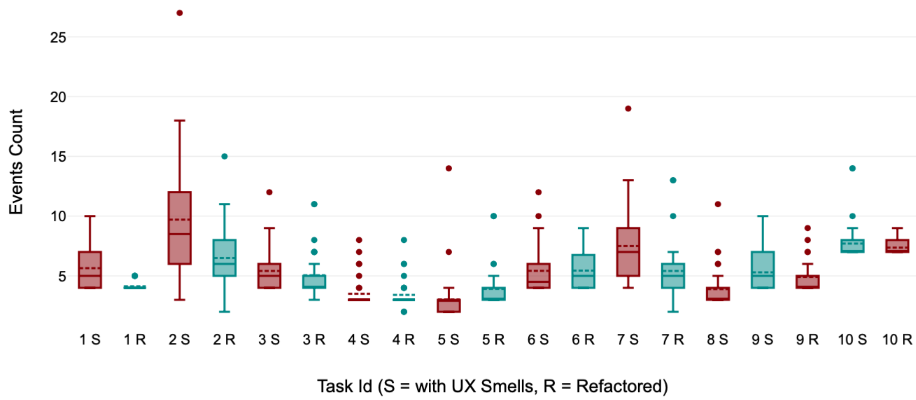

Fig. 4. Refactored and original (with smells) versions - Events per task.

for a keyword - Balcón Plantas) that took both more time and event count in the refactored version - and also rejecting H0 in favor of the smells version for the events count, as we explained in the previous subsection. Task 6 (Complete a subscription - Balcón Plantas) also showed higher numbers in both variables, but only slight differences (0.97 seconds / 0.02 events). The remaining case is task 9 (Click 3 "Read More..." links - Kabytes) for time, but also by a small margin - 0.21 seconds longer in average in the refactored version. Regarding the time variable, tasks 5 and 6 (i.e. using the search box and the subscription form in the "Balcón Plantas" website respectively) presented slightly higher times in the refactored version.

4.3 Threats to Validity

The experiment was subject to potential biases, regarding both internal and external validity.

With respect to internal validity, both circuits handed to the participants had the same order of applications. This could have impacted on the results, particularly the first and last application of each circuit. Even if this could have been managed from the start, we believe the impact was limited, since the 3 websites were essentially different, and the tasks were simple and did not require special preparation. These two aspects limit the impact of a potential learning effect.

External validity perhaps faced more potential threats due to the nature of the study. The first threat is the limited number of UI tested (3) which could have not been sufficiently representative. To address this, we selected popular websites that covered everyday tasks (like content search or simple forms) without depending too much on their domains. Completion bias could have played a significant role, i.e. participants could have been compelled to finish the tasks given to them, while in real settings they could have simply abandoned them if they got difficult. Even if we did not use effectiveness as a measure for this study (and all users were able to finish their tasks anyways), this effect could have also pushed the participants to finish earlier, which could have had an impact on efficiency measurement. However, this is also a very difficult threat to mitigate, since capturing (and especially delimiting) tasks *in the wild*, i.e. outside the context of an experiment, is a very challenging endeavor.

5 Post-Hoc Analysis on Events Bursts

To complement the study on event counts, we conducted a different analysis in which we grouped together events to form interaction "bursts" or units (Oulasvirta et al., 2005). This was motivated by our observations of the events timelines visualizations that we devised for the analysis of the evaluation with participants. Some of these visualizations evidenced small clusters of events that, intuitively, conformed small units of interaction, as can be seen in the sample timeline shown in Figure 5.

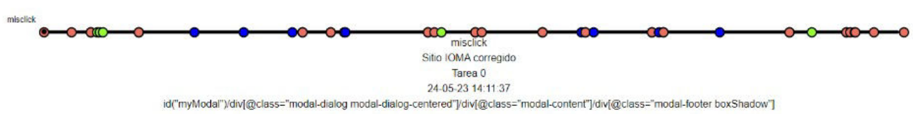

Fig. 5. Timeline visualization showing events' groupings.

Based on our observations of these timelines, we grouped the events on each session by different thresholds to separate related events, until we settled on 200 ms, which produced the groupings that best matched our observations, consistent with prior work on temporally bounded interaction units.

Table 3. Average Bursts per Session by task in refactored and original versions

Task	Site	UX Smell	With Smells	Refactored
1	IOMA	Too Close or Small Element	0.18	0.03
2	IOMA	Bad Readability	0.79	0.26
3	IOMA	Lost Search	0.50	0.47
4	Balcón plantas	Misplaced Menu	0.00	0.03
5	Balcón plantas	Lost Search	0.03	0.09
6	Balcón plantas	Unclear Form	0.16	0.24
7	Balcón plantas	Unclear Widget	1.16	0.97
8	Kabytes	Misplaced Menu	0.08	0.00
9	Kabytes	Unclear Widget	0.03	0.09
10	Kabytes	Lost Search	2.84	2.79

Table 4. Bursts: results of Mann Whitney tests for each task. Statistically significant (p< 0.05) in bold

Task	Site	UX Smell	p-value
1	IOMA	Too Close or Small Element	**0.03312**
2	IOMA	Bad Readability	**0.01069**
3	IOMA	Lost Search	0.47470
4	Balcón plantas	Misplaced Menu	0.86100
5	Balcón plantas	Lost Search	0.76320
6	Balcón plantas	Unclear Form	0.79700
7	Balcón plantas	Unclear Widget	**0.04260**
8	Kabytes	Misplaced Menu	**0.04992**
9	Kabytes	Unclear Widget	0.87610
10	Kabytes	Lost Search	0.41980

As shown in Table 3, sessions executed on versions containing UX smells tend to exhibit a higher average number of bursts than their refactored counterparts, although this pattern is not uniform across all tasks. Table 4 reports the corresponding statistical tests, indicating that a subset of tasks shows statistically significant differences ($p < 0.05$). In the remaining cases, while differences in average values are observed, they do not reach statistical significance. Taken together, these results indicate that burst-based measures are consistent with the trends observed in the event count and completion time analysis, and may serve as a useful complementary view.

6 Discussion

The results indicate that completion time and event count are influenced by the presence of smells. At first glance, there is a difference in most average values of these variables when considering sites with and without smells, with outliers confirming this trend. When the data were subjected to a hypothesis test, eight of the ten tasks passed in at least one of these two variables ($p < 0.05$).

In several cases where the test did not reject the null hypothesis on one of the two variables, there are indications that suggest that the failure may be more related to experimental design issues than to the predictive value of the presence of smells. An example in this regard is task 2, which failed the hypothesis test for time input. In this task, the user is instructed to find a text with a text layout that makes reading uncomfortable. In this case, the difficulty of the task, corroborated by the testimony of some participants, may have been the reason why several users took considerable time to solve it even in the corrected version. In other words, the challenge involved in a text interpretation exercise over several paragraphs may have acted as a confounding variable that ended up influencing the time taken as much or more than the presence of the smell in question.

In other cases, such as task 9 (activating "Read More..." buttons in the Kabytes website) the null hypothesis was not rejected for any of the variables.

The reason may be that in this case, the smell corresponded to a minor design flaw on the target buttons. Therefore, this design difference could have not represented a sizable influence on UX, so the lack of differentiation in the behavior of both versions was not surprising.

Certain smells that are expected to cause a significant difference in the number of events or in the time taken, but not in both at the same time. This is the case, for example, of tasks 4 and 10, which involve finding an element that is almost indistinguishable from the non-interactive content of the site.

Finally, from this results we could extract a list of the UX Smells that impact the most in terms of time and events count. Judging by the p-value, the top UX Smell to avoid would be **Too Small or Close Element**, both in time and events. Then, regarding events, the following two are **Unclear Widget** and **Bad Readability**, in that order. Regarding completion time, behind **Too Small or Close Element**, the remaining UX Smells are **Lost Search** and **Misplaced Menu**. However, it is worth noting that while the latter took consistently longer in both instances (tasks 4 and 8), the former only rejected H0 in one instance (task 10) and not in the other (task 3). This ranking of UX Smell severity can serve as a first pointer, but a more comprehensive experiment should be run that includes more instances of each UX Smell in different contexts.

7 Conclusions and Further Work

This article described an evaluation with real participants to assess the presence of UX Smells on the mobile web. The evaluation was focused on efficiency, as a particular aspect of the instrumental factors that compose the UX.

Results indicate that, in most cases, both completion times and event counts are higher when participants face UIs affected by known UX Smells that hinder their ability to complete the assigned tasks. In some of these cases (7 out of 20) these differences were statistically significant. Moreover, the results allowed prioritizing UX Smells in terms of severity.

The burst-based analysis is consistent with the trends observed for completion time and raw event counts. Tasks executed on versions containing UX smells generally exhibited a higher number of interaction bursts, suggesting that inefficient designs not only increase the amount of interaction, but also fragment user actions into shorter, fast-paced sequences. While this analysis is exploratory, it supports the interpretation of bursts as a complementary efficiency-related indicator rather than an independent construct.

Even if the evaluation is limited in the number of test websites, we still consider the tasks and the UX Smells representative of the current status of the mobile web. Nevertheless, more testing is required with a larger number of websites, smells and participants.

Currently, we are still analyzing the generated events through visualizations, which already motivated us to conduct the analysis on interaction bursts. We expect these timelines to help us find behavioral efficiency-related patterns that could entail new UX Smells not explored before. This analysis will also be useful for developing methods to detect these mobile UX smells automatically.

Regarding the UX smells impact on the users, we also plan to study the relationship of the smells with other UX factors beyond usability, specially with subjective aspects such as users preferences or perceptions. In this way, we can have a broader perspective about how each smell affects the overall user experience.

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

References

Ali, A., Xia, Y., Navid, Q., Khan, Z. A., Khan, J. A., Aldakheel, E. A., & Khafaga, D. (2024). Mobile-ui-repair: A deep learning based ui smell detection technique for mobile user interface. *PeerJ Computer Science*, *10*, e2028.

Almeida, D., Campos, J. C., Saraiva, J., & Silva, J. C. (2015). Towards a catalog of usability smells. *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, 175–181. https://doi.org/10.1145/2695664.2695670

Alonso-Virgós, L., Espada, J. P., Thomaschewski, J., & Crespo, R. G. (2020). Test usability guidelines and follow conventions. useful recommendations from web developers. *Computer Standards & Interfaces*, *70*, 103423.

Buono, P., Caivano, D., Costabile, M. F., Desolda, G., & Lanzilotti, R. (2020). Towards the detection of ux smells: The support of visualizations. *IEEE Access*, *8*, 6901–6914. https://doi.org/10.1109/ACCESS.2019.2961768

Burmistrov, I., Zlokazova, T., Izmalkova, A., & Leonova, A. (2015). Flat design vs traditional design: Comparative experimental study. In J. Abascal, S. Barbosa, M. Fetter, T. Gross, P. Palanque, & M. Winckler (Eds.), *Human-computer interaction – interact 2015* (pp. 106–114). Springer International Publishing.

Chen, H.-J., Wu, M.-D., Lo, I.-C., & and, D. T.-H. C. (2024). Exploring the affordance effect of visual properties associated with virtual keyboard buttons on smartphones. *International Journal of Human–Computer Interaction*, *0*(0), 1–20. https://doi.org/10.1080/10447318.2024.2389349

de Santana, V. F., & Baranauskas, M. C. C. (2015). Welfit: A remote evaluation tool for identifying web usage patterns through client-side logging. *International Journal of Human-Computer Studies*, *76*, 40–49. https://doi.org/https://doi.org/10.1016/j.ijhcs.2014.12.005

Gardey, J. C., Garrido, A., Firmenich, S., Grigera, J., & Rossi, G. (2020). Ux-painter: An approach to explore interaction fixes in the browser. *Proc. ACM Hum.-Comput. Interact.*, *4*(EICS). https://doi.org/10.1145/3397877

Garrido, A., Firmenich, S., Rossi, G., Grigera, J., Medina-Medina, N., & Harari, I. (2013). Personalized web accessibility using client-side refactoring. *IEEE Internet Computing*, *17*(4), 58–66. https://doi.org/10.1109/MIC.2012.143

Grigera, J., Garrido, A., Rivero, J. M., & Rossi, G. (2017). Automatic detection of usability smells in web applications. *International Journal of Human-Computer Studies*, *97*, 129–148. https://doi.org/https://doi.org/10.1016/j.ijhcs.2016.09.009

Harms, P., & Grabowski, J. (2014). Usage-based automatic detection of usability smells. *Human-Centered Software Engineering: 5th IFIP WG 13.2 International Conference, HCSE 2014, Paderborn, Germany, September 16-18, 2014. Proceedings 5*, 217–234.

Harms, P., Herbold, S., & Grabowski, J. (2014). Trace-based task tree generation. *Proceedings of the Seventh International Conference on Advances in Computer-Human Interactions (ACHI 2014). XPS-Xpert Publishing Services.*

Limaylla Lunarejo, M. I., Santos Neto, P. d. A. d., Avelino, G., & Britto, R. d. S. (2020). Automatic detection of usability smells in web applications running in mobile devices. *Proceedings of the XVI Brazilian Symposium on Information Systems*, 1–8.

Nebeling, M., Speicher, M., & Norrie, M. (2013). W3touch: Metrics-based web page adaptation for touch. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2311–2320.

Oulasvirta, A., Tamminen, S., Roto, V., & Kuorelahti, J. (2005). Interaction in 4-second bursts: The fragmented nature of attentional resources in mobile hci. *Proceedings of the SIGCHI conference on Human factors in computing systems*, 919–928.

Paternò, F., Schiavone, A. G., & Pitardi, P. (2016). Timelines for mobile web usability evaluation. *Proceedings of the International Working Conference on Advanced Visual Interfaces*, 88–91. https://doi.org/10.1145/2909132.2909272

Paternò, F., Schiavone, A. G., & Conti, A. (2017). Customizable automatic detection of bad usability smells in mobile accessed web applications. *Proceedings of the 19th international conference on human-computer interaction with mobile devices and services*, 1–11.

Weichbroth, P. (2020). Usability of mobile applications: A systematic literature study. *Ieee Access*, *8*, 55563–55577.

Yang, B., Xing, Z., Xia, X., Chen, C., Ye, D., & Li, S. (2021). Uis-hunter: Detecting ui design smells in android apps. *2021 IEEE/ACM 43rd International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)*, 89–92.