POLITECNICO DI TORINO

MASTER's Degree in COMPUTER ENGINEERING

MASTER's Degree Thesis

Evaluating a Privacy-Oriented Browser Extension in an Experimental Context

Supervisors

Prof. Antonio VETRÒ

Ph.D. Lorenzo LAUDADIO

Candidate

Shayan HAJIAN

JULY 2026

Evaluating a Privacy-Oriented Browser Extension in an Experimental Context

Shayan Hajian

Abstract

Web browsing includes many background requests that most users do not see or pay attention to. Many web pages communicate with advertising networks, analytics services, social media platforms, and other external organizations. Since these communications happen behind the browser interface, users often have a limited understanding of which companies and organizations receive information while they browse the web. This thesis evaluates the results of using AwEE in a controlled in-class experiment. AwEE is a browser extension that uses gamification mechanisms to highlight privacy-related activities on the web and make them easier for users to understand. Before doing the experiment, we reviewed AwEE and made several improvements to it. Earlier, it had two separate versions: one for Chrome-based browsers and the other for Mozilla Firefox. We merged them into a single shared codebase. We fixed some visual problems it had based on some previous comments we received. We also increased the response speed of AwEE's interface and expanded the data that is reported within each exported session report. We also added references and news cases to our documentation to provide better explanations of why certain hosts have been considered risky. After making these revisions to AwEE, we conducted a controlled in-classroom study at Politecnico di Torino. Participants first completed a pre-test questionnaire and then used AwEE during a controlled web-browsing activity. After that, they answered the post-test questions. The same ten core privacy-related questions were included in both the pre-test and post-test and were answered using a five-point Likert scale. The pre-test also included a yes-or-no question about previous experience with privacy tools. In contrast, the post-test included a question about willingness to use AwEE or similar tools in the future. For the post-test, students also answered the System Usability Scale (SUS) questions. Anonymous session summaries generated by the extension were also collected and analyzed together with the questionnaire responses. The comparison between the pre-test and post-test results showed statistically significant increases in Privacy Concern and Attitude Toward Privacy Tools. Privacy Awareness was already high before doing the activity. We saw smaller changes in Privacy Awareness, Perceived Control, and Cautious Online Behavior, which were not statistically significant. The usability results showed that AwEE was generally considered usable, and most participants said that they would be willing to use AwEE or a similar privacy tool in the future. The session reports also showed many background requests and helped demonstrate how AwEE can make hidden web interactions more visible to users. Overall, the results show that using AwEE in this short in-class experiment improved some privacy-related attitudes, especially concern about how personal data is used

and a more positive attitude toward privacy tools. However, these results are based on a small group of people and a short-term evaluation. More studies with larger and more diverse groups, and with longer periods of usage, are needed to understand whether these improvements remain over time and can also be observed during normal everyday browsing.

Acknowledgements

I would like to dedicate this thesis, first of all, to my soulmate, Niloofar. Thank you for your endless love, patience, and support. You have given me strength, motivation, and comfort throughout this journey, especially in the moments when I needed them the most.

To my beloved mother, Shahin, and my dear father, Hamid, words cannot fully express my gratitude for your endless love and support. You have always stood beside me, encouraged me, and believed in me, even when the path was difficult. Your love has been my foundation, your prayers and guidance have been my strength, and your sacrifices have made this achievement possible. I am deeply thankful for everything you have done for me.

To my brother Arsham, my sister Shadi, and my grandmother, thank you for your love, encouragement, and belief in me. Your support has always meant a lot to me.

To my cousins, Arash and Kaveh, thank you for being part of this journey in your own meaningful ways.

To my best friends, Amir and Mahmoud, thank you for always being there for me. Your friendship, loyalty, support, and sense of humor made even the most difficult moments easier to face.

I would also like to sincerely thank Professor Vetrò for giving me the opportunity to work on this thesis, and for his guidance and support throughout the process.

Finally, I would like to thank Lorenzo for his help, availability, and kindness. His support and willingness to help whenever needed were truly appreciated.

To all of you, thank you.

Table of Contents

Chapter 1

Introduction

1.1 Context and Motivation

Visiting a website does not simply mean downloading a webpage. A simple visit can lead to communications with data analytics platforms, advertising networks, embedded media providers, social media services, and other external systems. Some of these connections are vital for proper functioning of the web pages. However, some other connections have measurement, advertising, or personalizing purposes. Most of these activities happen in the background, and the web pages that the user sees do not include all of them.

In recent years, many technical activities have been done to make these connections more clear. But the issue still persists. Consent banners can decide what requests should be sent, and their effectiveness depends both on the choice of the user and the way a website implements consent [1]. Some of these operations can be moved to servers, and be less obvious in the users' browsers [2]. New studies about browser transparency show that showing the methodologies adopted to use the data, can help users in their browsing sessions to see the information that they would not see in a normal situation as it would remain hidden [3]. A technical report can contain a lot of useful information. But without having the needed background and knowledge, it can be difficult to understand for a normal user. Because of that, a privacy awareness tool should not just collect the requests. It should also show them to the user in an understandable way.

1.2 AwEE and the Research Problem

AwEE is a web browsing extension that was made to answer the need for more clarity in a web browsing session. This extension combines request visualization with gamification mechanisms at the same time [4]. In a web browsing session, this extension analyzes the requests that are made, and compares the target domains from the requests with the ones that can be found in a predefined blacklist. If that specific domain exists in that list, it considers that as a *Bad Host*. A request that is sent to a Bad Host is called a *Bad Request*. The term Bad Host means that the

destination that the web request is sent to is related to some documented concerns and public cases that we found relevant and important for the purpose of providing privacy awareness to the users.

AwEE shows the information of the web browsing session, explains why each chosen host is considered a Bad Host by showing the related documents to each of those hosts, and uses gamification mechanisms to keep the user engaged. Its purpose is to visualize some of the requests that are sent in the background and to motivate the users to check them and be aware of them. So it should be considered as an educational tool.

The first version of AwEE was already evaluated [4]. What we present in this thesis goes beyond that and does it in two main steps. First, we checked the extension code and improved it in several ways. Second, we validated the extension in a structured in-class activity.

1.3 Aim of the Thesis and Research Questions

We have three main goals in this thesis. First, we want to know that after using the extension, do participants report changes in the specific privacy-related constructs or not. Second, we want to know if AwEE is considered usable by the participants. Finally, we want to know what the session logs show about the number of requests, percentage of Bad Requests, and the most frequent companies in the sessions.

The study is guided by the following research questions:

RQ1 Do participants' scores in the measured privacy-related constructs change after using AwEE?

RQ2 Is AwEE considered usable after the controlled in-class experiment?

RQ3 What session-level request counts, Bad Request percentages, and company-ranking patterns do we see during the controlled activity?

The first question will be answered by analyzing the answers of the participants to the pre-experiment/post-experiment paired items. These items cover Privacy Awareness, Privacy Concern, Perceived Control, Cautious Online Behavior, and Attitude Toward Privacy Tools.

The second question will be answered by calculating the SUS (System Usability Scale) scores of the participants, which is a well-known standard for this purpose.

The third question will be answered by checking the JSON sessions exported from the AwEE extension itself.

1.4 Contributions

This thesis presents these contributions:

- An updated implementation of AwEE extension;

- A detailed explanation of the design of an in-class experiment that includes pre questions, post questions, SUS questions, and collected session logs;
- Analyzing the observed short-term changes in the five different privacy-related constructs based on the answers of the participants;
- Discussing the limitations of the study and possible future ways to improve the extension and its evaluation.

Chapter 2

Background and Related Work

In this chapter we present the main concepts needed to understand the evaluation of AwEE. Our focus is not only on web tracking itself. We also discuss how web tracking can be visible to users, how privacy tools can be evaluated from the usability point of view, and why gamification can be useful in a privacy-awareness activity.

2.1 Web Tracking and User Transparency

A web page can be in contact with many domains in addition to the domain that the user enters. These connections might be related to content delivery, analytics, advertising, authentication, embedded media, social widgets, or other purposes. Some of them are vital for the correct functionality of the web page, while some others can be used for measurement, advertising, or observing users' behavior on different websites even when it is not necessary. From the user's point of view, these requests are normally hidden behind the visible page.

Studies have shown that third-party tracking exists deeply in ordinary browsing [5, 6]. These works are older than some recent studies that we discuss in this thesis, but they are useful because they explain the technical basis of third-party requests and large-scale tracking measurement. They also explain why browser-side observation can be a meaningful starting point for transparency.

Jha et al. studied how web requests change when users make choices in consent banners, such as accepting or rejecting tracking [1]. They showed that the requests a website sends do not depend only on the user's decision and on what the website claims. They also depend on how the consent mechanism is implemented in that specific website. For example, two websites may react differently to the same request of the same user because their consent systems are implemented differently. This is relevant to the AwEE extension because it does not rely on what the website claims. It is capable of showing the real request activity in real time during a web browsing session, without trusting the cookie-banner claims and consent-interface statements.

At the same time, browser-based observation has its own limits. Fouad et al. studied server-side tracking, where part of the data processing can happen through external infrastructure that is less visible from the user's browser side [2]. Their

study showed that some tracking does not happen only inside the user's browser. Instead, part of it can be moved to external servers or hidden behind domains that look like they belong to the visited website. Because of this, a browser-based tool may not be able to see the full tracking process. It can show the requests that leave the browser, but it may not show what happens to the data after it reaches the server. This is one of the problems that AwEE faces. It can report the requests that are visible on the browser side, but it cannot reconstruct every possible data flow after a request reaches a server. Because of this reason, AwEE should be considered a tool to increase transparency and not a complete system to monitor users' privacy.

Zimmeck et al. investigated different ways to visualize website data transparency directly inside the browser [3]. They implemented a browser extension called Privacy Pioneer that included a privacy pop-up, a privacy history interface, and watchlist notifications. Their study showed that browser-integrated privacy interfaces can help users understand website data-collection practices. Most participants found the tool easy to understand and useful. This path is closer to the target of AwEE. Privacy-related information becomes more meaningful when it appears at the same time that the user is browsing the web, and when it is related to the current page they visit. This makes privacy information more connected to the user's real browsing activity. Instead of only reading a general privacy policy separated from the page, users can understand which background requests are sent while they visit a page and how these requests can be related to their privacy.

2.2 Browser Privacy Tools

Several browser extensions try to block, limit, or explain activities that are harmful to users' privacy. NoScript provides good control over scripts and active content for users, but its configuration requires technical knowledge [7]. Privacy Badger detects possible trackers and tries to limit them without looking at a pre-determined manual list [8]. uBlock Origin uses filter lists and detailed per-site controls to block advertising and tracking content [9]. FPMON focuses on browser fingerprinting and shows the detected activity to the user [10].

These tools have different purposes. Some of them prioritize protection through blocking, while others mostly focus on transparency and reporting. AwEE belongs to the second group and it can be considered a learning tool to increase users' awareness. This extension shows the list of web requests to users, links the risky ones to companies and related news cases, and uses gamification mechanisms to encourage users to investigate more.

2.3 System Usability

The usability of a privacy tool is very important for its effectiveness. A tool can present detailed technical information, but if users do not understand what it talks about, it will be considered a failure. Corner et al. reported usability problems of

privacy browser add-ons, including the difficulty of their configuration and system feedback [11]. This supports the decision of this thesis to evaluate AwEE not only through the number of requests, but also through users' ideas about ease of use, consistency, learnability, and confidence.

The SUS is a questionnaire that Brooke introduced for a fast and overall evaluation of the usability of a system [12]. The statements alternate between positive and negative wording and use a five-point agreement scale for answering each one. After transforming the points, the final score for each user becomes a number between 0 and 100. This is a standard score which becomes meaningful by comparing it with reference distributions and adjective ratings [13].

Research on the factor structure of SUS has identified usable and learnable components, while also confirming the practical value of the overall score [14]. In this thesis, the total score is the main standard for measuring usability. We also analyze item distributions, as they show strengths and weaknesses that might stay hidden when we calculate only an overall average.

2.4 Gamification for Privacy Awareness

Gamification is defined as the usage of game-like mechanics inside contexts that are not games themselves [15]. Reviews show that the effectiveness of gamification depends on the mechanics that are chosen, the target users, and the type of the activity [16]. Points, avatars, and visual effects do not guarantee learning alone.

The AwEE design is mostly based on the Octalysis framework [17]. AwEE includes immediate visual feedback, curiosity about hidden requests, a simple profile/avatar, and a local record to give users a sense of progress. The purpose is not to turn privacy into entertainment, but instead is to make the user aware of the information that is normally invisible and to find reasons to inspect it. This distinction is ethically important. A privacy tool should not manipulate the user into anxiety or wrong certainty. The gamified elements inside AwEE should support attention and thinking, and at the same time should not hide the limitations of the tool.

2.5 Previous AwEE Work and Research Gap

The first study of AwEE described the extension and compared its educational and gamified approach with currently existing privacy tools [4]. That work made the initial concept and collected some initial feedback. In particular, the previous evaluation showed that the interface was promising, but also showed that some parts of the extension needed improvements before performing a structured activity.

The research gap that this thesis focuses on is related to the evaluation of a gamified transparency tool as part of a controlled educational activity. The experiment we designed, which will be presented in the next chapters, answers three perspectives. They are attitude changes before and after using AwEE, perceived

usability, and observing what happened during the web browsing sessions. Therefore, our experiment design covers all of them together.

Chapter 3

The AwEE Extension and the Improvements

In this chapter we discuss the AwEE extension and the version of it that was used in our experiment. We first describe what this extension does, and then we explain the updates and changes we made to the extension before performing the in-class activity.

3.1 Purpose and Main Concepts

AwEE is a browser extension that observes the web requests that are made during a web browsing session. The main purpose of this extension is to show the user some parts of the browsing activity that are usually not visible.

When a user visits a website, the browser does not always communicate only with that website. It can also send requests to many other hosts. Some of these requests are needed for the correct working of the page. Some others can be related to advertising, analytics, embedded media, social media services, or other external systems.

AwEE checks these requests and compares their destination with a list that already exists in the project. This list contains hostnames related to some companies. For these companies, AwEE also keeps some references and public cases that explain why they are important for the purpose of privacy awareness.

In this thesis we use two main terms:

Bad Host A destination hostname that exists in the AwEE blacklist.

Bad Request A request that is sent to a Bad Host.

This distinction is important because one host can receive many requests during the same session. These terms only mean that the destination is related to a company or service that AwEE considers relevant for privacy awareness based on the documents stored inside the extension, and we should not consider them legal, moral, or security judgments about a company.

3.2 User Interaction

The AwEE pop-up shows the current browsing session by using counters and a request log. When a Bad Request is found, the related row in the log is shown differently and the numbers are updated.

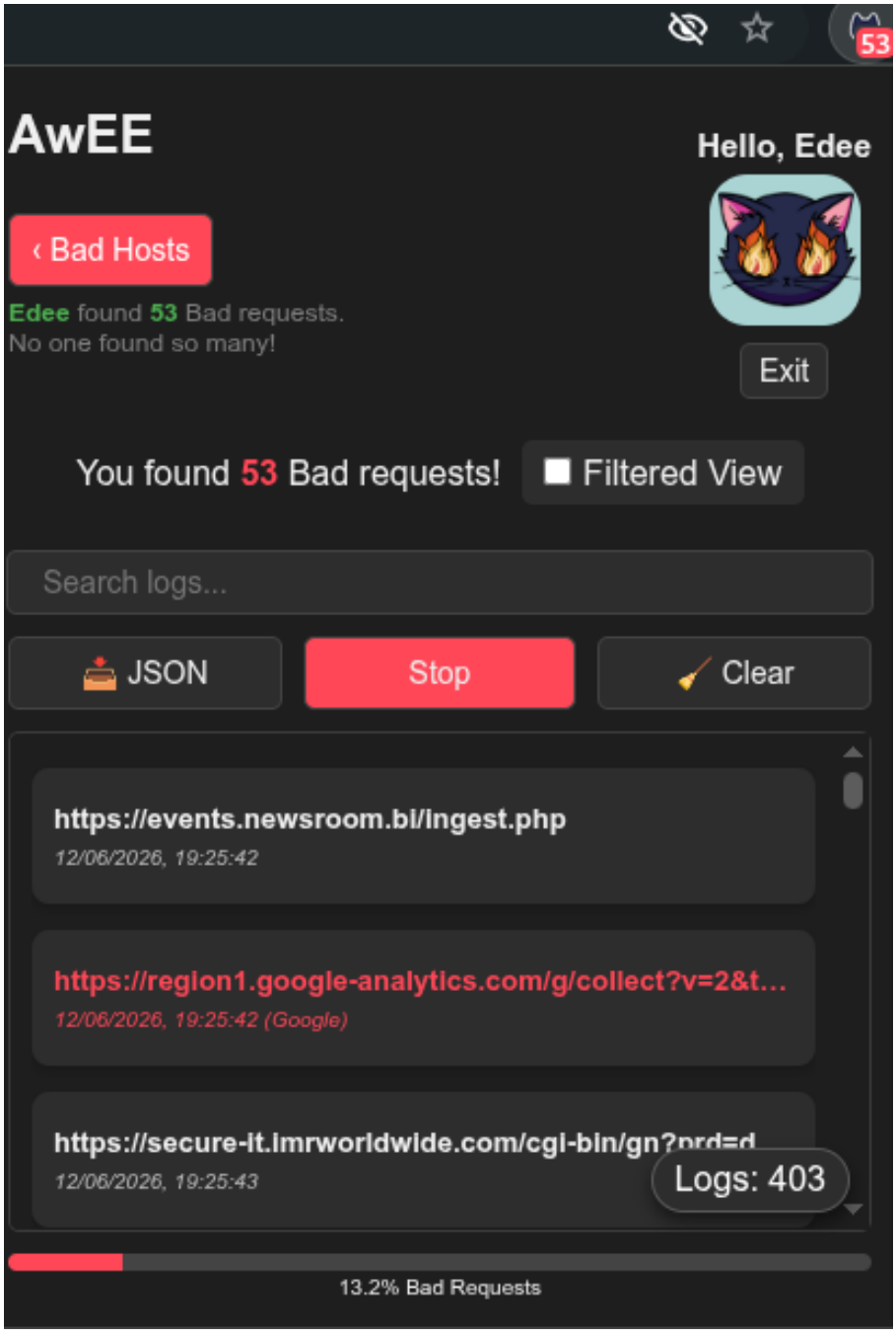


Figure 3.1: AwEE browser extension interface during a browsing session.

The user can check the destination of the request. The user can also open the references that are connected to that host. Because of that, the extension does not only show a number or a warning. It also gives the user a reason for why that host exists in the list.

AwEE also has a small cat avatar. The cat has different versions, and its expression changes when the number of Bad Requests increases. The extension also stores locally the highest number of Bad Requests observed in previous sessions.

These parts were added to give feedback to the user. They also give a simple sense of progress. The goal is not to turn AwEE into a complete game. The purpose is to motivate the user to pay attention to the requests and check them.

In the updated version used in this thesis, we also added a counter on the extension icon. This counter shows the current number of Bad Requests directly on the AwEE logo. This change was useful because the user can still see the main number even when the pop-up is closed.

3.3 Browser Support and Unified Architecture

In the earlier version of AwEE, the code for Firefox and Chromium-based browsers was separated. This means that some similar parts of the extension existed in two different codebases. This was not very good for maintaining the extension because if we fixed a problem in one version, there was still a chance that the same correction was not done in the other version.

Before doing the classroom study, we reorganized the code by using Mozilla's WebExtension browser API polyfill [18]. This polyfill helps to reduce some differences between browsers. For example, it helps with the way browsers use namespaces and asynchronous APIs.

After this change, the main logic of the extension became shared between the two browser versions. Some settings still had to remain different, because Firefox and Chromium-based browsers do not use exactly the same manifest settings. But the main part that records and processes the requests, and also the interface code, became common.

This change was useful for the experiment too. It made the extension easier to test and update. It also reduced the work needed to prepare the two versions that were distributed to the participants.

3.4 Request Recording and Responsiveness

The background logic of AwEE was also revised before doing the experiment. In longer sessions, we had observed that some request information could be lost or not handled correctly. In the previous implementation, when the request history became large, the pop-up could become slow. Sometimes it could also be temporarily not responsive.

In the updated version, we fixed this issue. Now, the complete session record is kept separated from the part that is currently shown in the interface. So now, when the user searches in the log or uses filters, only the displayed list changes. The stored session data does not change.

This was important for our study because each participant browsed for 35 minutes, and the pages used in the activity could generate thousands of requests. The exported report had to show the complete session, even if the user had searched or filtered the list during the activity.

Another problem was also fixed in the filtered view. In the previous interface, when the user selected the view that showed only Bad Requests, the percentage indicator showed 100% all the time. This happened because all the requests visible in that filtered list were Bad Requests. But for the user it was not useful because the percentage did not refer to the full session anymore. In the updated version, this percentage is hidden in the filtered view. But when the user returns to the normal view, it is shown again.

3.5 Exported Session Summary

At the end of a session, AwEE creates a JSON file. In the updated version used in this thesis, this JSON file was considered as the only required export format for the experiment. The base version also included the option to download the session log summary as a CSV file.

We did not use a separate CSV export. The reason was to avoid receiving different kinds of files from participants. If more than one format was possible, some participants could send the wrong file or a file that was not needed for the analysis.

In the updated version of the session summary, we added these fields that did not exist in the previous version:

- total number of requests;
- total number of Bad Requests;
- percentage of Bad Requests over all recorded requests;
- the three companies that had the highest number of Bad Requests;
- percentage of all Bad Requests related to each of those companies;
- session duration.

The reason we added those fields was to make the analysis easier and lighter by reducing the needed preprocessing and to make the JSON file more understandable outside of the extension. Because exported request information can include full request URLs, these files were treated as potentially sensitive research data.

3.6 Bad Host List and Supporting Evidence

We rechecked the Bad Host list before doing the experiment. Our reason was not only to add more companies to the list, but also to make the list more reasonable,

more consistent, and clearer for the user. This list is a project resource maintained by the project team.

A company was included in the updated version of the Bad Host list only if two conditions were satisfied. First, we had to find documented privacy-related concerns or public cases about that company. Second, we had to recheck all of the existing domains of the previous version to make sure that they belonged to that specific company.

The supporting documents could have different types. For example, privacy-policy material, watchdog or civil-society analyses, research papers, regulatory actions, lawsuits, investigations, fines, breaches, or other public controversies. So, adding a company required both a privacy-awareness reason and a domain check.

We did the domain checking very carefully. They were all checked using public sources such as company pages, registry information, ownership descriptions, and public documentation. One of the main sources that we used for this purpose was `who.is` [19]. We also did not keep the domains that we were not sure who owned. So, if the ownership was not clear, or if the privacy reason was not strong enough, the domain or company was removed from the list.

We also added Apple, TikTok, Temu, OpenAI, Xiaomi, and Huawei to the list. This decision was made by considering the feedback that was received from the previous AwEE evaluation. In that evaluation, some participants were surprised that these companies were not considered Bad Hosts [4].

At the same time, we removed five companies from the list. We did it because, after more research, we understood that the available documents were not strong enough for keeping them in this version. These companies were Intuition Machines, Fonticons, StackPath, AddToAny, and Fastly.

The supporting file that explains the hosts was also extended. In the base version, detailed descriptions and evidence existed only for a smaller group of companies. In the updated version, we also added the descriptions of the remaining companies too. Because of that, every Bad Host now has its own explanation.

The supporting material was organized into two groups:

References Background or explanatory materials. They can include privacy policies, criticism pages, watchdog analyses, or research papers. These links help explain why the company is considered important for privacy awareness.

News cases Real public cases. They can include fines, investigations, breaches, lawsuits, or other controversies. These links show cases that are connected to the company.

Broken links inside the reference interface were also corrected. This was important because one of the goals of AwEE is to let users inspect the reason for a Bad Host classification. If the link does not open, the extension becomes closer to a warning tool without enough explanation.

3.7 Interface Revisions Based on Previous Feedback

Several changes were made inside the interface of the AwEE extension based on the previous feedback that was received [4].

Some of the participants had mentioned that the strong red colour that was used for highlighting the Bad Hosts and counters made reading hard. To address that, we changed it to a less saturated red. We wanted to keep the warning feeling, but also make its readability better.

Another feedback point was about long HTTP links. Some participants said that showing the full request in the pop-up could confuse users who do not know what an HTTP request is. Because of that, in the updated interface, very long links are shortened in the visible log. If a link has more than 130 characters, only the first part is shown and the rest is not. If a link is shorter than 130 characters, we show it completely.

However, we kept the full text of the request URLs in the exported JSON file. So we only changed what the user sees in the pop-up.

These revisions were not evaluated one by one. The classroom study evaluates the updated version of AwEE as a whole. Because of that, we cannot say that the final usability score was caused by one specific change. But it is still important to explain these revisions, because they show which version of AwEE was actually used in the experiment.

Chapter 4

Study Design and Procedure

4.1 Evaluation Strategy

We used a within-participant pre-test/post-test structure for the study. Each participant answered the first questionnaire before doing the controlled activity. After using AwEE, they answered the related items again.

This design was suitable for our classroom study because each participant acted as their own reference point. It does not remove all possible alternative explanations. But it helps us reduce the effect of stable differences between participants.

We organized the evaluation using the GQM approach [20]. In this approach, we first define the main goal of the evaluation. Then we connect this goal to research questions and measurable indicators. Table 4.1 shows the GQM goal definition that we used for evaluating AwEE.

Table 4.1: GQM goal definition for the AwEE evaluation.

Analyze	The AwEE browser extension and the sessions recorded with it.
For the purpose of	Evaluating short-term privacy-related outcomes, perceived usability, and session-level request summaries.
With respect to	The five privacy-related constructs, SUS usability, Bad Request exposure, and recurrent companies.
From the viewpoint of	End users and researchers.
In the context of	A controlled classroom activity in which participants visited seven predetermined websites.

4.2 Research Questions and Metrics

Table 4.2 connects each research question to its main evidence.

4.3 Participants and Sampling

Our participants were students from a university course at Politecnico di Torino. There were around 40 students in the introductory class, which was a week before the

Table 4.2: Research questions and primary metrics.

RQ	Question	Metrics
RQ1	Do participants' scores on the measured privacy-related constructs change after the AwEE activity?	Paired construct means, participant-level directions of change, item-level differences, and future-use intention.
RQ2	Is AwEE perceived as usable?	SUS total score, score distribution, and response distribution for the ten SUS items.
RQ3	What session-level request counts, Bad Request percentages, and company-ranking patterns are observed during the controlled activity?	Total requests, Bad Requests, Bad Request percentage, minimum and maximum percentage, protocol duration, and company rankings.

actual date of the experiment, while some other students followed the class remotely.

Participation was voluntary. We informed the students that the activity was part of the evaluation of the extension. We also explained that taking part in the activity was voluntary, and that the thesis would not report their personal information in any way.

For each participant, in order to be counted as a member of the experiment, we only considered those people who had completed a pre-test and a post-test, and had uploaded their AwEE session file together. If someone missed any of the requirements, we removed them from the experiment. At the end, nineteen participants met all of these requirements.

The participants were part of an academic context. Although the sample size was limited, it was manageable for giving technical support and collecting the files. But we should also consider the fact that it might not represent the general population. The literature about pilot studies shows that small sample sizes can still be useful for checking feasibility and getting preliminary estimates. But the results should be interpreted carefully [21]. So we believe this sample size is suitable for our study.

4.4 Timeline

The activity was divided into an introductory moment and a formal controlled session.

On 28 April 2026, we introduced AwEE to the students during a class. We showed them where they could download the extension from the university teaching portal. After that, we showed the installation and startup process step by step.

After we executed the extension, we opened a website as an example to show the students the functionality of the extension and how it works. We explained to them the counters, the request list, the meaning of Bad Host, and the references that are shown for the related hosts.

In the same class, we also presented the pre-test questionnaire to them. We mentioned that they could answer it during the class or later. The deadline to answer the pre-test questionnaire was 4 May 2026 at 23:59.

The reason that we organized an in-class introductory session before performing

the main experiment was that we did not want the experiment day to be mostly about installation problems. This helped the students become familiar with the extension before the main session. This decision also had one limitation. Some learning could already have happened before the formal controlled activity, because participants had already seen the extension, its counters, the request list, and the meaning of Bad Host. So the pre-test should be considered as a baseline before the formal controlled browsing activity, not as a baseline before any exposure to AwEE.

On 5 May 2026, we performed the experiment. Only the students who had completed the pre-test were included in this session.

At the beginning of the class, we explained the procedure again to the participants. We also reminded them that the collected data would be used only for research purposes, and that the results would be reported in aggregate form.

The browsing task started at 10:10 and lasted 35 minutes. During this time, participants visited seven websites. They spent five minutes on each website. Every five minutes, the instructor asked the class to continue with the next website.

After finishing the browsing task, participants had to complete the post-test questionnaire and the SUS questionnaire. They also had to export and submit their AwEE session file. The deadline for submitting the post-test, SUS questionnaire, and exported session file was 13:00 on the same day.

4.5 Selected Websites

We selected the websites before the activity. The reason for choosing these websites was that they were expected to be familiar for students in an Italian university context. They also helped us to see the different levels of third-party requests and Bad Requests during a common browsing task.

Before doing the activity, we checked these websites ourselves. Based on this check, we selected the order of the websites in a way that participants would progressively visit pages with more third-party requests or Bad Requests. Because of this, the list was not randomized.

The websites used in the activity were:

1. `wikipedia.org`, a widely used online encyclopedia [22];
2. `funzionepubblica.gov.it`, the Italian public-administration portal [23];
3. `polito.it`, the institutional website of Politecnico di Torino [24];
4. `subito.it`, an Italian marketplace for second-hand goods [25];
5. `reddit.com`, a large discussion and social platform [26];
6. `ilpost.it`, an Italian online newspaper [27];
7. `ilfattoquotidiano.it`, another Italian news website [28].

This list was only a controlled set that we used to create the same browsing task for all of the participants, and it helped us compare the results between participants more easily.

4.6 Questionnaire Design

We used a five-point Likert scale for the main attitude items. The scale was from 1, which means Strongly disagree, to 5, which means Strongly agree. Likert-type scales are mostly used to measure attitudes and the level of agreement of participants [29].

The pre-test questionnaire included ten Likert items and one Yes/No item. The Yes/No item asked them about the previous use of privacy tools.

In the post-test questionnaire, we repeated the first ten statements with the same wording. We did this because we wanted to compare the answers before and after the activity directly. If we had changed the wording, the difference in the answers could be caused by the phrasing and not by the activity itself, and could lead to bias.

We also added one separate item in the post-test questionnaire. This item measured whether participants were willing to use AwEE or a similar tool in the future or not. Participants also answered the ten SUS items in the post-test questionnaire in addition to the statements.

We have five constructs, and each construct has two items. So, in general, we have ten paired items. Table 4.3 shows the whole structure.

Table 4.3: Mapping of paired questionnaire items to privacy-related constructs.

Construct	Pre-test items	Post-test items
Privacy Awareness	PRE1, PRE2	POST1, POST2
Privacy Concern	PRE3, PRE4	POST3, POST4
Perceived Control	PRE5, PRE6	POST5, POST6
Cautious Online Behavior	PRE7, PRE8	POST7, POST8
Attitude Toward Privacy Tools	PRE9, PRE10	POST9, POST10

You can see the exact list of the questions in Appendix A. PRE11 and POST11 are not a pair, as they measure two completely different things. PRE11 asks about earlier experience with privacy tools, while POST11 asks about future willingness to use AwEE or a similar tool.

4.7 Collected Data and Participant Handling

For our analysis, we combined three sources of data:

- pre-test questionnaire responses;
- post-test responses, including SUS;
- the AwEE JSON summary from the controlled session.

The questionnaires and the JSON session files were submitted through the Politecnico di Torino teaching portal. This allowed us to match the pre-test, post-test, and session file that belonged to the same participant. After this matching step, we carried out the analysis using study records and aggregate results. The thesis reports only aggregate findings, not individual identities or individual browsing traces.

Only the project team had access to the collected data used for the evaluation. The session files were treated as potentially sensitive research data, because the exported request URLs may contain detailed browsing information. We did not analyze or report the exact URLs of individual participants. In this thesis, we only use these files to calculate aggregate values such as the number of requests, the percentage of Bad Requests, company rankings, and session duration.

This combination gave us both subjective and objective information. The questionnaires show the attitudes and perceptions of the participants. The session files show the requests that AwEE observed during the controlled task. We did not consider either of these sources as a complete measure of real privacy protection.

Chapter 5

Data Preparation and Analysis Method

5.1 Dataset Preparation

The final dataset included 19 records, each belonging to one participant, and each made of 40 columns. Each record included the pre-test questionnaire, post-test questionnaire, SUS questionnaire, and exported session log.

As mentioned earlier, all of the Likert responses were within the range from 1 to 5. Only students who submitted all of the needed data were considered in the experiment, and those who did not were not included in the dataset for analysis.

5.2 Construct Scores

For every privacy-related construct, we had two questionnaire items. Before comparing the pre-test and post-test answers, first we changed these two answers into one construct score. We did this by calculating the average of the two items for each participant.

$$C_i = \frac{x_{i1} + x_{i2}}{2}. \tag{5.1}$$

We did the same calculation for the pre-test and for the post-test. After that, for each participant, we calculated the difference between the post-test score and the pre-test score.

$$\Delta C_i = C_{i,\text{post}} - C_{i,\text{pre}}. \tag{5.2}$$

If the value was positive, it meant that the score increased after the activity. If the value was negative, it meant that the score decreased. If the value was zero, it meant that there was no change.

We used the construct scores to answer RQ1. We also analyzed the single items, in order to help us understand the construct results better. This way, we could see whether the two items inside the same construct changed in the same direction or

not.

5.3 Descriptive Statistics

Our analysis included the pre-test mean, the post-test mean, and the mean paired difference for every construct. We also calculated the number of participants whose scores increased, decreased, or stayed unchanged. The mean paired difference was used as the main standard for checking the size of the changes, because it is explainable on the same 1–5 scale that we used for the questionnaire.

The distribution of the Likert answers was also checked. Averages might have the same mean, but the answers of the participants can still be distributed in different ways. The percentage of agreement was calculated by combining the "Agree" and "Strongly Agree" answers together. This was done to help us explain the results better in some parts of our analysis. It was not a replacement for the original five-point answers.

5.4 Paired Statistical Tests

For the pre-test and post-test comparisons, we used the Wilcoxon signed-rank test. We used this test because the data were paired. Each participant had one answer before the activity and one answer after the activity.

This test was suitable for our case because it does not require the data to follow a normal distribution in the same way as a paired-samples t-test.

We applied the test to the five construct pairs. We also applied it to the ten individual item pairs, but only as a supporting analysis.

We chose $\alpha = 0.05$ as the significance threshold. The table related to the results of the constructs also reports the original p-values and the Holm-adjusted p-values for the five construct-level tests [30].

Holm adjustment was used because these five constructs were all about the comparison between pre-test and post-test, but they were tested separately. The item-level tests were reported to better understand the construct results, and they were explained more carefully.

This reporting procedure separates the main findings from the supporting analysis. The construct-level results directly answer the research question of the experiment. The item-level results explain which statements of the questionnaire affected the construct-level changes, and in which way. But they should not be considered as separate findings. They should be considered as supporting information.

5.5 SUS Scoring

We calculated the SUS score using the standard method [12]. For the positive items, we subtracted one point from the score. For the negative items, we subtracted the

score from five. At the end, we summed all the adjusted values and multiplied the result by 2.5.

$$\text{SUS} = 2.5 \left[\sum_{j \in \{1,3,5,7,9\}} (r_j - 1) + \sum_{j \in \{2,4,6,8,10\}} (5 - r_j) \right]. \tag{5.3}$$

The final SUS score is a value between 0 and 100. In addition to the average, we also included the minimum, maximum, standard deviation, and the number of participants who had a score equal to or higher than 70.

The 70-point threshold was only used as a practical reference, not as a universal pass or fail standard. The interpretation of SUS scores depends on reference distributions and adjective ratings [13].

5.6 Session Data Analysis

As we mentioned in the previous parts, the session of the AwEE extension included different parts, such as total request count, total Bad Request count, Bad Request percentage over total requests, top three companies with the highest number of Bad Requests, and the session duration.

For the company analysis, we checked which company appeared in each of the first three positions in the session summaries. This ranking is based on the AwEE list and also on the websites that we selected for the experiment.

Because of this, it should not be interpreted as a general ranking of companies across the whole web. It only shows the companies that appeared more in our controlled activity.

The exported sessions also include the session-duration field. But we did not consider it in our analysis as a key factor. The reason is that the experiment had a fixed time of 35 minutes for everyone.

5.7 Analysis Scripts, Reproducibility, and Reporting Precision

We implemented the quantitative analysis using a set of Python scripts. We did not include the scripts line by line in the main text, because the thesis is not code documentation. But their role is still important, because they define how the tables and figures were produced. Appendix B summarizes the main scripts and explains the purpose of each of them.

The flow of the analysis was as follows: first, we analyzed the construct scores and item scores. Then we computed the SUS scores, summarized the company rankings, and reviewed the final tables and figures.

We reported the numerical values with enough precision and we were careful to make the analysis traceable.

We also separated the measured findings from the interpretation. In the Results chapter, we report the values that we observed. In the Discussion chapter, we explain the possible reasons behind those results.

Chapter 6

Results

6.1 Data Quality and Participant Background

All of the 19 records included the needed data about the questionnaire and the session data. Sixteen participants, or in other words about 84% of them, mentioned prior use of privacy tools, and about 16% mentioned no previous experience. So most participants already knew about privacy tools before the study. This helps us explain why the first Privacy Awareness score was already very high.

6.2 RQ1: Construct Changes

Table 6.1 reports the comparison between five constructs. The biggest changes were observed in Privacy Concern and Attitude Toward Privacy Tools. The p column shows the original Wilcoxon p-value. The p_{Holm} column reports the Holm-adjusted value for each of the five constructs.

Table 6.1: Pre-test and post-test construct results ($N = 19$).

Construct	Pre	Post	Δ	Inc.	Dec.	Same	p	p_{Holm}
Privacy Awareness	4.447	4.553	0.105	4	1	14	0.1573	0.3146
Privacy Concern	3.447	3.842	0.395	10	1	8	0.0061	0.0305
Perceived Control	2.605	2.368	-0.237	4	9	6	0.0981	0.2943
Cautious Online Behavior	3.500	3.632	0.132	8	4	7	0.2836	0.3146
Attitude Toward Privacy Tools	3.526	4.053	0.526	13	1	5	0.0065	0.0305

Privacy Concern increased by 0.395 points, and went up from 3.447 to 3.842. The score of ten participants increased, for one person it decreased, and for eight of them it remained unchanged. The Wilcoxon result was statistically significant both before and after Holm correction ($p = 0.0061$, $p_{\text{Holm}} = 0.0305$).

Attitude Toward Privacy Tools had the biggest increase in the average. It was 0.526 points. The score of thirteen participants increased, for one person decreased, and for five participants stayed unchanged. This comparison was also statistically significant after Holm correction ($p = 0.0065$, $p_{\text{Holm}} = 0.0305$).

Figure 6.1: Mean pre-test and post-test scores for the five privacy-related constructs.

6.2.1 Item-level comparisons

Table 6.2 shows the results of our study on the ten paired items. The p-values in this table are not adjusted and should be considered only as supporting evidence.

The noticeable changes, even in the item-level comparison, also occurred for the items belonging to the same two constructs whose change was also significant at the construct level. If we apply Holm correction to all ten item pairs, none of the item-level results will remain significant at $\alpha = 0.05$.

Table 6.2: Item-level paired comparisons with unadjusted Wilcoxon p-values.

Pair	Pre	Post	Δ	Inc.	Dec.	Same	p
1	4.526	4.579	0.053	3	2	14	0.6547
2	4.368	4.526	0.158	3	0	16	0.0833
3	3.368	3.789	0.421	8	1	10	0.0209
4	3.526	3.895	0.368	8	1	10	0.0196
5	2.263	1.895	-0.368	2	7	10	0.0702
6	2.947	2.842	-0.105	3	4	12	0.5271
7	3.368	3.474	0.105	5	4	10	0.5637
8	3.632	3.789	0.158	4	1	14	0.1797
9	3.421	4.000	0.579	11	1	7	0.0209
10	3.632	4.105	0.474	9	1	9	0.0126

The agreement with the two statements of the Privacy Concern construct increased by 26.3 and 21.1 percentage points. The agreement with the statement that claimed that privacy tools are necessary increased by 36.8 points, and the agreement with the statement that claimed that these tools can reduce online tracking also increased by 31.6 points. See Figure 6.2.

Figures 6.3 and 6.4 visualize the complete distribution of the answers. These figures are useful to show the movement toward agreement in the Privacy Concern and Attitude Toward Privacy Tools items. The full text of all the statements, or items, is reported in Appendix A. In the figures, for better readability, we used the item codes instead of writing the text of each item.

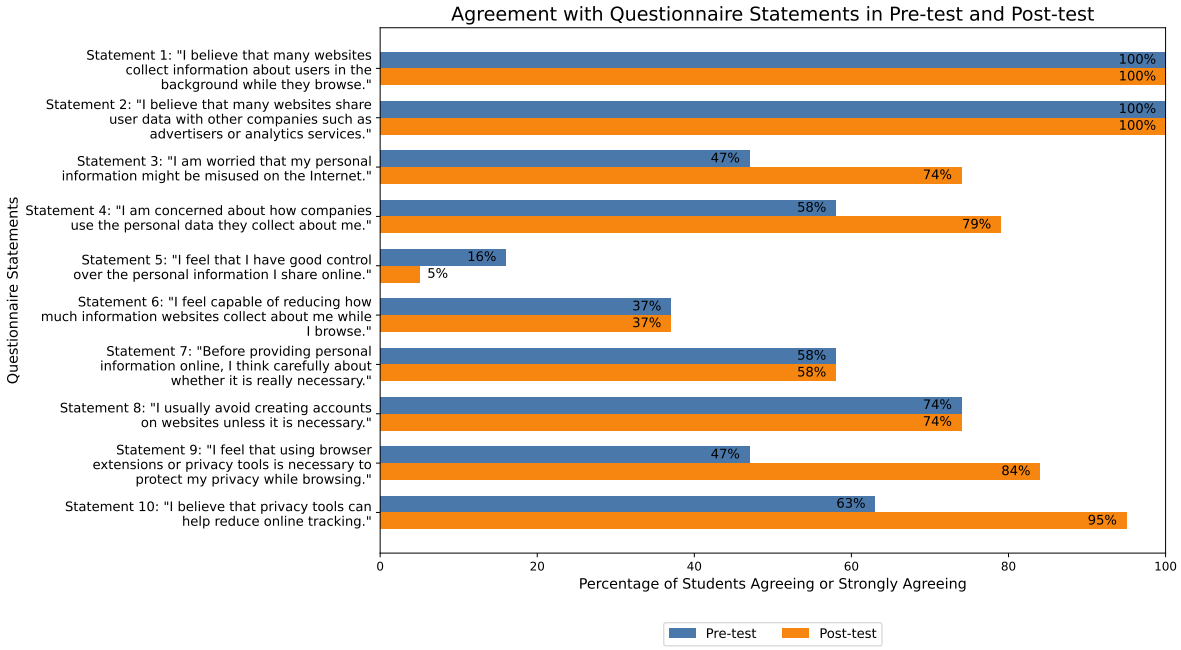


Figure 6.2: Agreement or strong agreement for selected pre-test and post-test items. The full wording of the questionnaire items is reported in Appendix A.

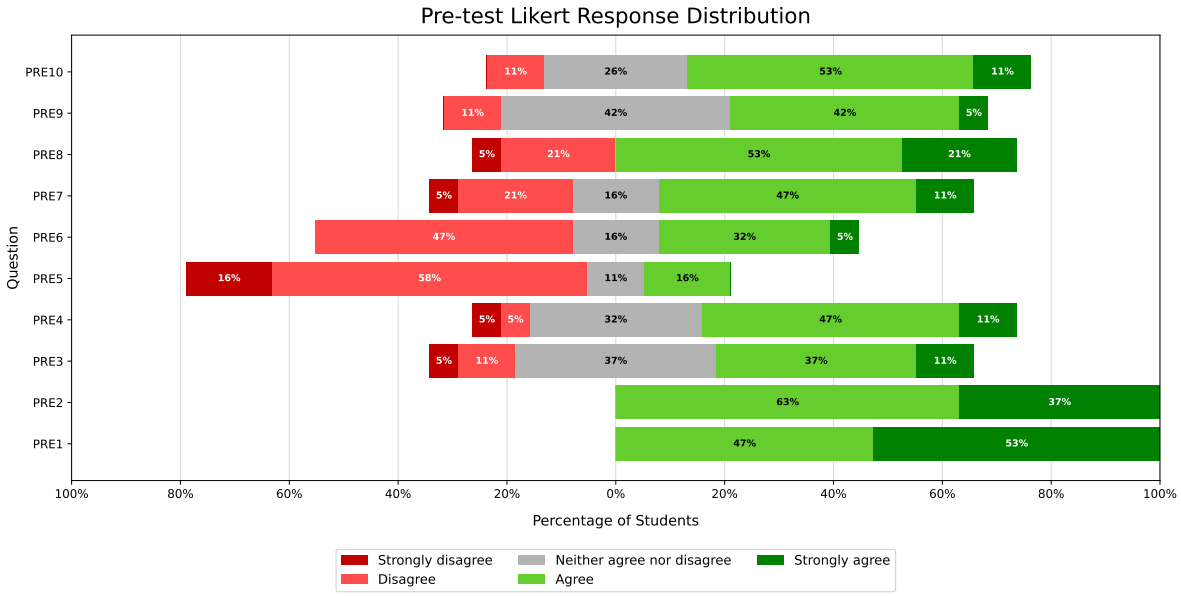


Figure 6.3: Pre-test Likert response distribution for PRE1–PRE10. The full item wording is reported in Appendix A.

6.2.2 Future willingness

Statement POST11 asked the participants if they were willing to use AwEE or a similar privacy tool in the future or not. Fourteen participants, or in other words about 74%, agreed or strongly agreed with this statement. One participant, which means about 5%, was neutral. Four participants, meaning about 21%, disagreed or strongly disagreed.

The detailed answers were one Strongly disagree, three Disagree, one Neutral, ten Agree, and four Strongly agree.

6.3 RQ2: Perceived Usability

AwEE achieved an average SUS score of 72.8 and a median score of 72.5. The scores were between 52.5 and 95, and the standard deviation was 10.57. Twelve participants had a SUS score of 70 or more, and seven participants had a score lower than 70.

Table 6.3: Summary of SUS results.

Measure	**Value**
Participants	19
Mean	72.8
Median	72.5
Standard deviation	10.6
Minimum	52.5
Maximum	95
Scores ≥ 70	12
Scores < 70	7

The item distribution shows that ease of use was the clearest strength of AwEE, because all of the participants either agreed or strongly agreed that the extension was easy to use. In addition, about 90% of them considered the functions of the extension well integrated. Also, about 90% believed that most people would learn to use it quickly, and about 74% of them mentioned that they felt confident when they were using it.

Regarding the items with negative meaning, it is important to notice that for these items, the more participants disagree, the better the score is, and disagreement for these items is a positive thing. About 90% disagreed that AwEE was unnecessarily complex. Also, about 90% disagreed that technical support would be necessary for using it. About 74% disagreed with the statement regarding inconsistency, and about 79% of the participants disagreed with the need for a lot of initial learning before using AwEE.

The weakest item was the one about interest in frequent use of AwEE. Only about 32% of the participants either agreed or strongly agreed that they would like to use AwEE frequently. This comparison shows an important difference. This statement showed us that most of the participants do not like to use AwEE regularly. But we

also had another statement, POST11, where about 74% of them mentioned that they would like to use AwEE in the future. This shows that most of the participants are more willing to use this kind of tool from time to time, rather than frequently.

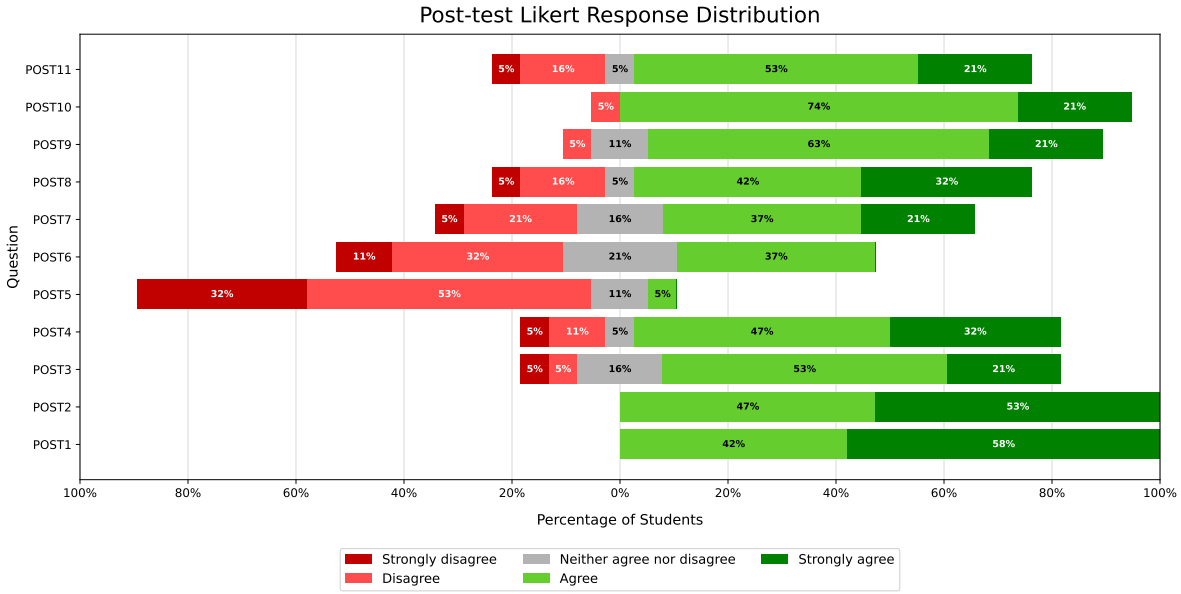
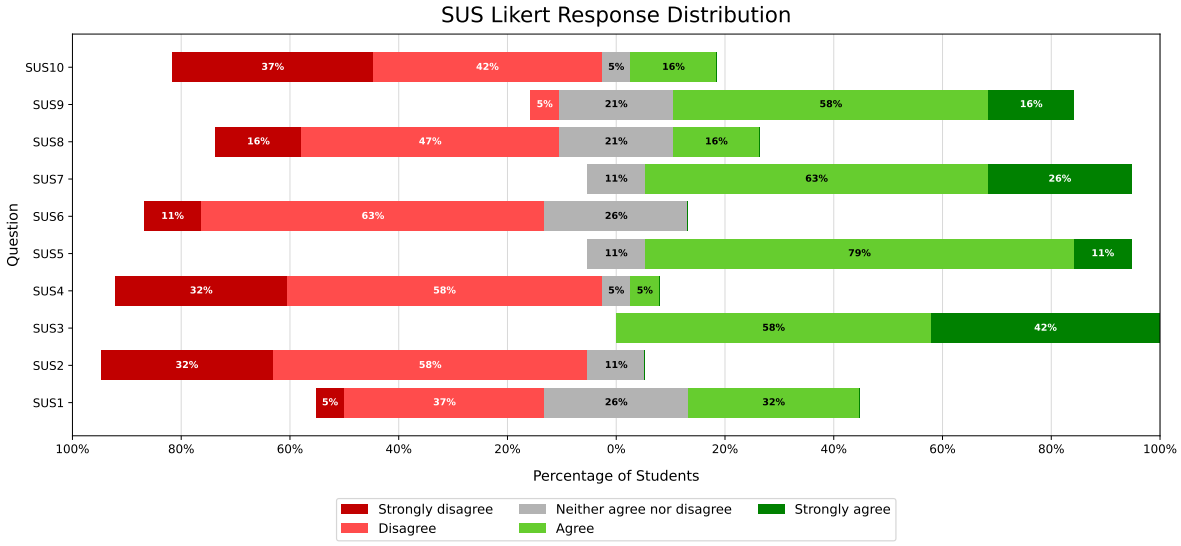


Figure 6.4: Post-test Likert response distribution for POST1–POST11. The full item wording is reported in Appendix A.



Note: SUS items 2, 4, 6, 8, and 10 are negatively worded. meaning that disagreement with these items reflects more positive usability perceptions.

Figure 6.5: SUS response distribution. Items 2, 4, 6, 8, and 10 are negatively worded.

6.4 RQ3: Session-Log Summaries

The exported session logs showed us that during the in-class experiment, on average, each participant had about 6865 total requests, and about 793 of them were Bad Requests. So, in other words, about 12% of all the requests were Bad Requests. The lowest percentage was 7% and the highest was 23%.

Table 6.4: Summary of AwEE session data.

Measure	Value
Participants	19
Protocol browsing duration	35 minutes
Mean total requests	6864.8
Mean Bad Requests	792.8
Mean Bad Request percentage	12.4%
Minimum Bad Request percentage	7.0%
Maximum Bad Request percentage	23.0%

These differences in values are expected and normal. It is true that all of the participants visited the same list of websites, but the exact number of requests still can vary. Caching, browser version, other installed extensions on the browser of the participants at the same time as AwEE, page updates, and some other small differences in the way that each participant interacted with the websites can affect the final number of requests.

The session log files also included the list of the top three companies with the highest number of Bad Requests. For all of the 19 participants, Google was always in the first place. Meta was the most common company in the second place, appearing about 53% of the time there. TikTok was also the most common company for the third place, appearing there about 47% of the time.

Table 6.5: Most common company at each exported ranking position.

Position	Company	Participants	Percentage
First	Google	19	100.0%
Second	Meta	10	52.6%
Third	TikTok	9	47.4%

If we consider the second and third positions together, for 15 participants, meaning 79% of them, Meta appeared in one of these two positions, and for TikTok, this number was 14 participants, which means about 74% of the time. This means that Google, Meta, and TikTok were the most common companies associated with Bad Requests in our experiment. These results are only about our in-class experiment, and are also related to the selected list of websites. Because of this, they should not be considered as a general ranking for all companies around the world.

Chapter 7

Discussion

7.1 Answer to RQ1

The clearest result for RQ1 was related to two constructs: Privacy Concern and Attitude Toward Privacy Tools. After the activity, the scores of the participants increased for both of them. These changes were statistically significant. They also remained significant after applying Holm correction to the five construct-level tests.

The item-level results also showed the same direction. Participants reported more concern about the misuse of their personal information and about the way companies collect and use data. They also agreed more with the idea that privacy tools are necessary and that they can reduce online tracking.

For Privacy Awareness, we did not see a significant increase. But we should not understand this as meaning that the activity had no educational value. The reason is that the initial Privacy Awareness score was already very high. Also, 16 out of 19 participants had already used privacy tools before. Because of this, there was not much space for this score to increase more. This is a ceiling effect. If we repeat the same activity with users who have less previous knowledge, the result may be different.

For Perceived Control, the score decreased a little. One possible explanation is that after seeing thousands of requests and repeated third-party companies, participants understood better how complex online tracking is. This can make users feel less in control, even if they become more concerned about privacy. However, this result was not statistically significant. So we cannot confirm this explanation with our data.

Cautious Online Behavior only changed a little. We should know that some behaviors, like avoiding creating unnecessary accounts or thinking before sharing online information, might need more time and repeated usage of the tool. As the total duration of our study was 35 minutes, and the participants were asked to answer the questions in a short period after doing the in-class activity, it is important to mention again that our goal was to measure self-reported answers, not real long-term behavior.

Overall, the answer to RQ1 is this: we observed significant short-term increases in Privacy Concern and Attitude Toward Privacy Tools. But we did not see statistically

significant changes in Privacy Awareness, Perceived Control, or Cautious Online Behavior.

7.2 Answer to RQ2

The average SUS score was 72.8, and it shows that, in general, participants considered AwEE usable. The item-level results show that the participants found AwEE easy to use, easy to learn, and well integrated. These results are important for a learning privacy tool, because if the tool is not usable enough, users may spend their attention on understanding the tool itself rather than focusing enough on the privacy-related information that AwEE shows. So in this way, the focus on learning would be weak. But we saw that the feedback was positive in general.

The intention to use AwEE regularly was weaker than the other SUS statements. At the same time, about three quarters of the participants mentioned that they are willing to use AwEE or another similar tool in the future. These two statements are not in conflict with each other. They do not show opposite results. A participant might like to use AwEE in some specific situations, but does not want to use it all the time during normal daily web browsing.

Future versions of AwEE should try to find a way to increase the willingness for regular usage.

7.3 Answer to RQ3

The session files showed that the controlled activity generated many background requests. On average, 12.4% of the recorded requests were Bad Requests. The values were different between participants, from 7.0% to 23.0%.

These numbers help us understand why a real-time tool like AwEE can be useful in an educational activity. When users visit a website, they usually only see the visible page. They do not normally see how many requests are happening in the background. AwEE makes this hidden activity visible, and because of that, users can understand the size of this activity better.

In the company rankings, Google appeared in the first position for all participants. Meta and TikTok were also common in the next positions. But this result should be read only in the context of our experiment. It depends on the selected websites, the browsing date, browser behavior, and the host list used by AwEE.

Because of this, these rankings should not be considered as a general statement about the whole web. They only describe what happened in our controlled classroom activity.

7.4 Implications for AwEE

Based on the results of this study, we can say these points about AwEE:

- The basic interaction of AwEE was understandable enough for a short classroom activity.
- Showing the requests in real time helped participants think more about privacy concerns and privacy tools.
- The number of Bad Requests alone is not enough. Users also need explanations and possible next steps.
- Participants found AwEE easy to use and easy to learn, but this did not mean that they wanted to use it frequently.
- The educational value of AwEE should not be judged only based on the number of Bad Requests.

7.5 Threats to Validity

In this section, we explain the main threats to the validity of our study. We organized them based on the common categories used in empirical software-engineering research [31].

7.5.1 Construct Validity

For each of the five privacy-related constructs, we used two custom items. We discussed these items and mapped them through the GQM approach, but we did not take them from a previously validated multi-item scale.

Because of this, the items may not cover all parts of privacy awareness, privacy concern, perceived control, cautious online behavior, or attitude toward privacy tools. Also, the answers were self-reported. So, we should remember that some participants may have answered in a way that they thought was more expected or more acceptable.

Using the same wording supports paired comparison because we wanted to compare the answers directly. But there can be a limitation. Participants might remember their older answer to a statement, and therefore they might be willing to answer the same statement the same way again. For usability, this problem is less strong because we used SUS, which is an established questionnaire. However, we changed the word *system* to *extension*, and this is still an adaptation.

Another construct choice is the Bad Host list. This list is based on the criteria of the project and on the documented cases that we considered relevant. It is not an objective legal classification of every request. Because of this, the number of Bad Requests depends on this list. So even after the recheck and updating of the Bad Host list, based on any methodology, this list is still a project resource and should not be considered a legal assessment of all web activity. Also, it is important to mention that AwEE's blacklist does not include all of the third parties that might be related to privacy issues.

7.5.2 Internal Validity

We did not have a randomized control group in our experiment. The observed changes between pre-test and post-test might be affected by the explanation of the instructor and the context of the class course. Participants were all from the Data Ethics and Data Protection course. Also, as explained earlier, we had an introductory session before performing the formal experiment. So there might be some possibility that some kind of learning could also happen before the experiment.

All of the participants visited the same websites in the same order. This helped us compare the results between participants. But it also creates an order effect. The websites were selected and ordered on purpose. The reason was that we wanted participants to gradually visit pages that generated more third-party requests or Bad Requests. If the order of the websites was different, the impression of the participants could also be different.

Also, the differences in browser cache, browser version, other installed extensions on the participants' systems running at the same time as AwEE, network conditions, and dynamic page content could affect the final session results as well.

7.5.3 Conclusion Validity

The final sample had only 19 participants. The statistical power is limited because of that. The results can be affected by individual participants.

We also performed several construct-level and item-level tests. To avoid making the results look stronger than they are, we reported Holm-adjusted p-values for the construct-level results. The item-level results were treated only as supporting results.

The item-level results are useful because they help us to understand the construct results in more detail. But we should not consider them as separate confirmations. After applying Holm correction across the ten item-level tests, none of the individual item-level results was significant at $\alpha = 0.05$. Because of this, in this thesis, we focused mainly on the construct-level findings.

7.5.4 External Validity

The participants were master's students in a university course related to data ethics and privacy. Most of them had already used privacy tools before. Because of this, the results may not be the same for users with different ages, education, technical knowledge, or motivation.

The activity also lasted only 35 minutes and included seven predefined websites. Normal daily browsing is different from this. It is more personal, more varied, and usually happens for a longer time. Because of this, the findings of this study should be understood as the results of a controlled short-term learning activity. They should not be considered as evidence of continuous real-world use or long-term behavior change.

7.6 Ethical and Interpretive Caution

Participation in the activity was voluntary. We told the students that their personal information would not be reported in the thesis. The collected data were used only for the evaluation, and only the project team had access to them.

We matched the questionnaires and the session files through the university's teaching portal. We asked the participants to submit their results there. This thesis only reports the aggregate results, not the individual identities of the participants in any form.

We should also note that if we use privacy tools without explaining their limitations, it can create the wrong feeling of certainty. AwEE checks the client-side requests and matches them with its own list. But it is important to know that it cannot show every server-side operation, and it cannot decide the legal meaning of each data flow.

Because of this, we use careful language in this thesis. Our results do not show complete privacy protection, and they also do not prove harmful behavior by every company included in the Bad Host list.

Chapter 8

Conclusion and Future Work

This thesis evaluated AwEE as a gamified browser extension to study web privacy awareness in a classroom activity. The work included two main parts. First, we updated the extension. Then, we evaluated it with participants in a controlled activity.

Our in-class experiment consisted of a pre-test, a controlled 35-minute web browsing task, a post-test that also included SUS questions, and the session summaries exported by AwEE. Nineteen participants uploaded all of the required data for our analysis, and only the participants who provided the full material were included in the analysis.

After analyzing the collected data, the strongest result that we observed was related to Privacy Concern and Attitude Toward Privacy Tools. After using AwEE, the scores of the participants increased for these two constructs. These two results were still significant after applying Holm correction to the five construct-level tests.

AwEE was generally considered usable by the participants. The average SUS score was 72.8. The strongest results at the item level were related to ease of use, integration, and ease of learning. Most of the participants also mentioned that they were willing to use AwEE or another similar tool in the future. However, most of them were not interested in using AwEE frequently.

To sum up, the results show that AwEE was a useful tool for a short-term privacy-learning activity in this controlled classroom context. But we should not consider these results as proof of long-term behavior change, or a general result for all types of users. The conclusions of this thesis are limited by the small sample of participants, the fact that the participants were all from the same university class, the fixed order of websites, and the fact that the post-test was done not long after the browsing session. Also, the introduction that was presented one week before the experiment date could have affected the results.

AwEE can be improved in some different directions in the future.

First of all, after diagnosing a Bad Request, it can provide some suggestions to the user. The user might understand that a third party exists in the web browsing session, but still might not know what action to take after that. Some short descriptions can be helpful to talk about browser privacy settings, consent choices, reliable blocking

tools, and also the limits of each of these actions, and suggest to the user what to do. It can also explain the difference between requests that are needed for the normal working of the page and requests that are more related to advertising, analytics, profiling, or social-media integrations. This can help users understand not only that a Bad Request happened, but also what kind of web activity it may represent.

Secondly, AwEE can switch between a learning mode and another daily standard mode. The learning mode can provide more details, explanations, and include more gamification mechanisms. But the everyday mode, on the other hand, can include fewer gamification mechanisms, and only summarize the web browsing activity from time to time for the user. This might increase the willingness for regular usage of AwEE, without making it annoying or asking for too much attention.

Third, the Bad Host list and the supporting documents for each of the Bad Hosts already exist. But an option can be added to the extension to give the user the opportunity to remove some domains from the list, or add some others that they consider risky for any personal reason. In addition, we can add a description as a banner, or in any other form, to clearly mention that AwEE is not a legal, moral, or security judgment about all of the companies or every request related to them.

Fourth, future versions can also test more gamification mechanisms. But this should be done carefully. Adding more points, badges, or competition does not mean that users will learn more about privacy for sure. The gamification elements should only be added if they help the user understand the privacy-related activity better. When possible, the effect of each mechanism can also be evaluated separately.

A future study should include more participants, and also should include participants with different levels of technical knowledge and privacy experience. This would strengthen causal interpretation. Randomizing or counterbalancing website order would reduce sequence effects.

Longer observation is also necessary. A follow-up questionnaire after several weeks could also test whether concern, tool attitudes, or behavior remain different. With the permission of the participants, future work could also analyze real privacy-setting changes instead of only relying on immediate self-reports.

Future studies can also include some open-ended questions, as they might be able to explain some kind of results that the numerical analysis cannot. By doing that, we might be able to better understand why frequent-use intention was low for some participants, why Perceived Control tended to decrease, and which explanations of AwEE were most useful for the participants.

The present thesis also used two-item construct averages and item-level inspection. In a future study, there can be more items for each construct, and also the statements can be refined before using them for a broader evaluation with more participants.

The current study is a starting point for this work: the extension was usable enough for a structured activity, the evaluation protocol can be reproduced, and the observed short-term changes justify the need for a bigger and more accurate follow-up study.

Appendix A

Questionnaires

A.1 Custom Pre-test Items

PRE1–PRE10 used the following scale: 1 = Strongly disagree, 2 = Disagree, 3 = Neither agree nor disagree, 4 = Agree, 5 = Strongly agree.

PRE1. I believe that many websites collect information about users in the background while they browse the web.

PRE2. I believe that many websites share user data with other companies such as advertisers or analytics services.

PRE3. I am worried that my personal information might be misused on the Internet.

PRE4. I am concerned about how companies use the personal data they collect about me.

PRE5. I feel that I have good control over the personal information I share online.

PRE6. I feel capable of reducing how much information websites collect about me while I browse.

PRE7. Before providing personal information online, I think carefully about whether it is really necessary.

PRE8. I usually avoid creating accounts on websites unless it is necessary.

PRE9. I feel that using browser extensions or privacy tools is necessary to protect my privacy while browsing.

PRE10. I believe that privacy tools can help reduce online tracking.

PRE11. In the past, I have used additional privacy tools, such as privacy extensions, ad blockers, or anti-tracking tools. (Yes/No)

A.2 Custom Post-test Items

POST1–POST11 used the same five-point agreement scale.

POST1. I believe that many websites collect information about users in the background while they browse the web.

POST2. I believe that many websites share user data with other companies such as advertisers or analytics services.

POST3. I am worried that my personal information might be misused on the Internet.

POST4. I am concerned about how companies use the personal data they collect about me.

POST5. I feel that I have good control over the personal information I share online.

POST6. I feel capable of reducing how much information websites collect about me while I browse.

POST7. Before providing personal information online, I think carefully about whether it is really necessary.

POST8. I usually avoid creating accounts on websites unless it is necessary.

POST9. I feel that using browser extensions or privacy tools is necessary to protect my privacy while browsing.

POST10. I believe that privacy tools can help reduce online tracking.

POST11. I am willing to use this extension or similar tools in the future to protect my privacy.

A.3 System Usability Scale Items

The standard SUS items were adapted only by referring to “this extension” instead of “the system”. All items used the five-point agreement scale.

SUS1. I think that I would like to use this extension frequently.

SUS2. I found this extension unnecessarily complex.

SUS3. I thought this extension was easy to use.

SUS4. I think that I would need the support of a technical person to be able to use this extension.

SUS5. I found the various functions in this extension were well integrated.

SUS6. I thought there was too much inconsistency in this extension.

SUS7. I would imagine that most people would learn to use this extension very quickly.

SUS8. I found this extension very cumbersome to use.

SUS9. I felt very confident using this extension.

SUS10. I needed to learn a lot of things before I could get going with this extension.

Appendix B

Supplementary Result Tables

B.1 Agreement Changes for Selected Items

Table B.1: Participants answering Agree or Strongly agree.

Item summary	Pre	Post	Change
Concern about misuse of personal information	47.4%	73.7%	+26.3 points
Concern about companies' use of personal data	57.9%	79.0%	+21.1 points
Privacy tools are necessary	47.4%	84.2%	+36.8 points
Privacy tools can reduce tracking	63.2%	94.7%	+31.6 points

B.2 SUS Item Contributions

Each adjusted SUS item contributes between 0 and 4 points before multiplication by 2.5.

Table B.2: Mean raw responses and adjusted SUS contributions.

Item	Item type	Raw mean	Contribution
SUS1	Positive	2.842	1.842
SUS2	Negative	1.789	3.211
SUS3	Positive	4.421	3.421
SUS4	Negative	1.842	3.158
SUS5	Positive	4.000	3.000
SUS6	Negative	2.158	2.842
SUS7	Positive	4.158	3.158
SUS8	Negative	2.368	2.632
SUS9	Positive	3.842	2.842
SUS10	Negative	2.000	3.000

B.3 Company Ranking Detail

For the second ranking position, Meta appeared for 10 participants, TikTok for 5, and Amazon for 2. For the third position, TikTok appeared for 9 participants, Meta for 5, Apple for 2, X/Twitter for 2, and Cloudflare for 1. These values are descriptive and depend on the fixed browsing task and the AwEE list used on the activity date.

B.4 Multiple-testing Note

The five construct-level tests are the main inferential results of the thesis. Their Holm-adjusted p-values are reported in Chapter 6. The item-level tests were used to explain the construct-level findings. When Holm correction is applied across the ten item-level tests, none of the item-level results remains significant at $\alpha = 0.05$. Therefore, the item-level results should be interpreted as descriptive support rather than as separate confirmatory findings.

B.5 Analysis Scripts

The code of our analysis used for producing the results, tables and figures, is available on Zenodo at this address: `https://doi.org/10.5281/zenodo.20555128`. Table B.3 summarizes the purpose of each script.

Table B.3: Summary of the analysis scripts.

File	Purpose
`utils.py`	Defines shared paths, data-loading functions, questionnaire column groups, construct mapping, output folders, and the standard SUS scoring function.
`F1_data_check.py`	Checks dataset shape, column names, duplicate participant identifiers, missing values, Likert ranges, prior privacy-tool experience, session summaries, and company frequencies.
`F2_pre_post_construct_analysis.py`	Computes construct scores, participant-level construct changes, item-level changes, Wilcoxon signed-rank tests, and the pre/post construct figure.
`F3_sus_analysis.py`	Calculates SUS scores, SUS summary statistics, item contribution values, and SUS-related output tables and figures.
`F4_company_ranking_analysis.py`	Summarizes the companies appearing in the first, second, and third ranking positions in the exported session files.

File	Purpose
`F6_likert_plots.py`	Builds Likert count and percentage tables and creates the pre-test, post-test, and SUS Likert distribution figures.
`F7_final_summary_tables.py`	Produces final compact tables for participant overview, construct results, SUS, future willingness, and company ranking.
`F8_agreement_increase_chart.py`	Calculates agreement percentages for selected pre/post items and creates the agreement increase figure.
`F9_future_willingness_chart.py`	Summarizes POST11 willingness responses and session exposure values.
`F10_review_outputs.py`	Checks that expected tables and figures were produced and prints key results for final verification.

The scripts were used to generate the tables and figures reported in the thesis. The main text describes the methods rather than the implementation line by line, while this appendix documents how the analysis code was organized.

Bibliography

[1] Nikhil Jha, Martino Trevisan, Luca Vassio, and Marco Mellia. "The Internet with Privacy Policies: Measuring the Web Upon Consent". In: *ACM Transactions on the Web* 16.3 (2022), pp. 1–24. DOI: 10.1145/3555352 (cit. on pp. 1, 4).

[2] Imane Fouad, Cristiana Santos, and Pierre Laperdrix. "The Devil is in the Details: Detection, Measurement and Lawfulness of Server-Side Tracking on the Web". In: *Proceedings on Privacy Enhancing Technologies* 2024.4 (2024), pp. 450–465. DOI: 10.56553/popets-2024-0125 (cit. on pp. 1, 4).

[3] Sebastian Zimmeck, Daniel Goldelman, Owen Kaplan, Logan Brown, Justin Casler, Judeley Jean-Charles, Joe Champeau, and Hamza Harkous. "Website Data Transparency in the Browser". In: *Proceedings on Privacy Enhancing Technologies* 2024.2 (2024), pp. 211–234. DOI: 10.56553/popets-2024-0048 (cit. on pp. 1, 5).

[4] Luca Calò, Lorenzo Laudadio, Antonio Vetrò, Riccardo Coppola, and Marco Torchiano. "AwEE – Improving Privacy Awareness by Leveraging Gamification Principles". In: *Proceedings of the 2025 International Conference on Information Technology for Social Good*. GoodIT '25. ACM, 2025, pp. 411–419. DOI: 10.1145/3748699.3749819 (cit. on pp. 1, 2, 6, 12, 13).

[5] Franziska Roesner, Tadayoshi Kohno, and David Wetherall. "Detecting and Defending Against Third-Party Tracking on the Web". In: *Proceedings of the 9th USENIX Symposium on Networked Systems Design and Implementation*. NSDI '12. USENIX Association, 2012, pp. 155–168. URL: https://www.usenix.org/conference/nsdi12/technical-sessions/presentation/roesner (cit. on p. 4).

[6] Steven Englehardt and Arvind Narayanan. "Online Tracking: A 1-million-site Measurement and Analysis". In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. CCS '16. ACM, 2016, pp. 1388–1401. DOI: 10.1145/2976749.2978313 (cit. on p. 4).

[7] Giorgio Maone. *NoScript: Block Scripts and Own Your Browser!* 2025. URL: https://noscript.net/ (cit. on p. 5).

[8] Electronic Frontier Foundation. *Privacy Badger*. 2025. URL: https://privacybadger.org/ (cit. on p. 5).

[9] Raymond Hill. *uBlock Origin—Free, Open-Source Ad Content Blocker*. 2025. URL: https://github.com/gorhill/uBlock (cit. on p. 5).

[10] Julian Fietkau, Kashyap Thimmaraju, Felix Kybranz, Sebastian Neef, and Jean-Pierre Seifert. "The Elephant in the Background: A Quantitative Approach to Empower Users Against Web Browser Fingerprinting". In: *Proceedings of the 20th Workshop on Privacy in the Electronic Society*. Virtual Event, Republic of Korea: ACM, 2021, pp. 167–180. ISBN: 978-1-4503-8527-5. DOI: 10.1145/3463676.3485599. URL: https://dl.acm.org/doi/10.1145/3463676.3485599 (cit. on p. 5).

[11] Matthew Corner, Huseyin Dogan, Alexios Mylonas, and Francis Djabri. "A Usability Evaluation of Privacy Add-ons for Web Browsers". In: *Design, User Experience, and Usability. Practice and Case Studies*. Vol. 11586. Lecture Notes in Computer Science. Springer, 2019, pp. 442–458. DOI: 10.1007/978-3-030-23535-2_33 (cit. on p. 6).

[12] John Brooke. "SUS: A "Quick and Dirty" Usability Scale". In: *Usability Evaluation in Industry*. Ed. by Patrick W. Jordan, Bruce Thomas, Ian Lyall McClelland, and Bernard Weerdmeester. Taylor and Francis, 1996, pp. 189–194. DOI: 10.1201/9781498710411-35 (cit. on pp. 6, 20).

[13] Aaron Bangor, Philip Kortum, and James Miller. "Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale". In: *Journal of Usability Studies* 4.3 (2009), pp. 114–123. URL: https://dl.acm.org/doi/10.5555/2835587.2835589 (cit. on pp. 6, 21).

[14] James R. Lewis and Jeff Sauro. "The Factor Structure of the System Usability Scale". In: *Human Centered Design*. Ed. by Masaaki Kurosu. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 94–103. ISBN: 978-3-642-02806-9. DOI: 10.1007/978-3-642-02806-9_12 (cit. on p. 6).

[15] Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart E. Nacke. "From Game Design Elements to Gamefulness: Defining "Gamification"". In: *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*. MindTrek '11. ACM, 2011, pp. 9–15. DOI: 10.1145/2181037.2181040 (cit. on p. 6).

[16] Juho Hamari, Jonna Koivisto, and Harri Sarsa. "Does Gamification Work? A Literature Review of Empirical Studies on Gamification". In: *Proceedings of the 47th Hawaii International Conference on System Sciences*. IEEE, 2014, pp. 3025–3034. DOI: 10.1109/HICSS.2014.377 (cit. on p. 6).

[17] Yu-kai Chou. *Actionable Gamification: Beyond Points, Badges, and Leaderboards*. Packt Publishing, 2019. ISBN: 9781839210778. URL: https://books.google.com/books?id=9ZfBDwAAQBAJ (cit. on p. 6).

[18] Mozilla. *WebExtension browser API Polyfill*. 2026. URL: https://github.com/mozilla/webextension-polyfill (cit. on p. 10).

[19] Who.is. *WHOIS Domain Lookup*. 2026. URL: https://who.is/whois (cit. on p. 12).

[20] Victor R. Basili, Gianluigi Caldiera, and H. Dieter Rombach. "The Goal Question Metric Approach". In: *Encyclopedia of Software Engineering*. Ed. by John J. Marciniak. Vol. 1. John Wiley & Sons, 1994, pp. 528–532. URL: https://www.cs.umd.edu/~basili/publications/technical/T89.pdf (cit. on p. 14).

[21] Melody A. Hertzog. "Considerations in Determining Sample Size for Pilot Studies". In: *Research in Nursing & Health* 31.2 (2008), pp. 180–191. DOI: 10.1002/nur.20247 (cit. on p. 15).

[22] Wikimedia Foundation. *Wikipedia*. 2026. URL: https://wikipedia.org/ (cit. on p. 16).

[23] Presidenza del Consiglio dei ministri. *Ministro per la Pubblica Amministrazione*. 2026. URL: https://funzionepubblica.gov.it/ (cit. on p. 16).

[24] Politecnico di Torino. *Politecnico di Torino*. 2026. URL: https://www.polito.it/ (cit. on p. 16).

[25] Subito.it. *Subito*. 2026. URL: https://subito.it/ (cit. on p. 16).

[26] Reddit Inc. *Reddit*. 2026. URL: https://reddit.com/ (cit. on p. 16).

[27] Il Post. *Il Post*. 2026. URL: https://ilpost.it/ (cit. on p. 16).

[28] SEIF S.p.A. *Il Fatto Quotidiano*. 2026. URL: https://ilfattoquotidiano.it/ (cit. on p. 16).

[29] Rensis Likert. "A Technique for the Measurement of Attitudes". In: *Archives of Psychology* 140 (1932), pp. 1–55. URL: https://books.google.com/books?id=9rotAAAAYAAJ (cit. on p. 17).

[30] Sture Holm. "A Simple Sequentially Rejective Multiple Test Procedure". In: *Scandinavian Journal of Statistics* 6.2 (1979), pp. 65–70. URL: https://www.jstor.org/stable/4615733 (cit. on p. 20).

[31] Robert Feldt and Ana Magazinius. "Validity Threats in Empirical Software Engineering Research—An Initial Survey". In: *SEKE*. 2010, pp. 374–379. URL: https://www.cse.chalmers.se/~feldt/publications/feldt_2010_validity_threats_in_ese_initial_survey.pdf (cit. on p. 32).