



On the investigation of Usability in the Caixa Tem Application

José Vieira   [Universidade Federal do Agreste de Pernambuco | josevieira1709@gmail.com]

Rodrigo Andrade  [Universidade Federal do Agreste de Pernambuco | rodrigo.andrade@ufape.edu.br]

 Universidade Federal do Agreste de Pernambuco, Av. Bom Pastor, s/n, Boa Vista, Garanhuns, PE, 55292-270, Brazil.

Received: 30 November 2023 • Accepted: 22 April 2024 • Published: 27 May 2024

Abstract There is a growing advancement of technologies, with an increasing number of services being carried out through them. This trend also prompts the Public Sector to join this movement by providing its services digitally, making them more convenient and cost-effective. However, the challenge arises in ensuring that everyone has equal access to these services, thereby avoiding segregation, particularly among the most vulnerable segment of society. Hence, we investigate the usability of the Caixa Tem application, which provided access to emergency aid during the pandemic. Our goal is to verify whether this application ensures usability for its users. Therefore, we assess usability through high-fidelity prototypes faithful to the real application, conducting Heuristic Evaluations, and Usability Tests with carefully selected individuals. Therefore, we are able to identify a number of usability issues regarding time and interaction efficiency, satisfaction, user experience, and effectiveness. At last, we propose a set of modifications that enhances the usability of the Caixa Tem application.

Keywords: Usability Testing, Heuristic Evaluation, System Design and Usability, Public Systems, Caixa Tem.

1 Introduction

We live in a time when the use of technology is notoriously in various situations of our daily lives. There are numerous applications of Information and Communication Technologies (ICTs) to solve or facilitate tasks that we commonly perform [Kosakowski, 1998]. In this context, we have changed our way of communicating, interacting, and consuming products and services, for example. As a result, we increasingly encounter new technological solutions and applications that emerge to simplify and make our daily lives more convenient. This imminent contemporary digital culture ends up inducing the market to increasingly embrace and adapt to the digital realm, taking advantage of this medium to market their products and services. Therefore, it is now common for companies that are not present in the digital space to be losing market share [Gere, 2009].

In this context, gradually, the Brazilian Public Administration has also been taking advantage of this cultural shift, and as a result, it has been incorporating technology into the services provided to the population. This enables the development of a public management environment capable of keeping pace with these changes and also benefiting from the advantages generated by technology [Marchionini *et al.*, 2003]. This became evident during the COVID-19 pandemic [Lisbôa *et al.*, 2021; Monteiro *et al.*, 2022] when, due to the necessary social distancing, technology was chosen as a means to make the Brazilian Emergency Aid (Article 2° of Law No. 13,982/2020) accessible to the Brazilian society. To achieve this, the Brazilian Public Administration, in partnership with Caixa Econômica Federal¹, launched the Caixa Tem app on April 6, 2020, as a solution for social services and banking transactions, with the purpose of facilitating access to Emergency Aid for the population [Caixa Econômica Federal -

CEF, 2020].

For the intended purpose of Caixa Tem, it is important that the application maintains good quality, ensuring good usability for users who need its services [Sahasrabudhe and Lockley, 2014; Monteiro *et al.*, 2023; Filho *et al.*, 2023]. Otherwise, the system may cause difficulty in accessing essential services for the Brazilian population. In this way, a problem of segregation of the population may arise regarding the services provided by the Brazilian Public Administration. A portion of the population benefits from these services because they can use Caixa Tem, while another portion is disadvantaged by not being able to use it [Viana, 2020]. Therefore, the main objective of our study is to investigate the usability of the Caixa Tem application. Thus, we aim to identify usability issues and propose modifications that correct and prevent the problem from recurring.

To conduct this usability investigation of Caixa Tem, initially, we develop a high-fidelity prototype [Budde *et al.*, 1990] of the application using the Figma [Figma, 2011] tool, aiming to make it as faithful as possible to the original Caixa Tem app. Subsequently, we conducted a usability evaluation using Heuristic Evaluations [Nielsen, 2005]. These evaluations lead to a series of improvements for the identified usability issues, which we implement for a new prototype version. Subsequently, we proceed to evaluate the usability of the app again through Usability Tests [Lewis, 2006]. In these tests, we assess our two prototype versions (i.e., original and after fixes). We use five metrics to measure the usability of each prototype: time efficiency, interaction efficiency, effectiveness, user experience, and satisfaction. After these initial tests, we identify further improvements that could be made to the application. Thus, we create a new version of the prototype. Finally, we conduct new Usability Tests with this updated prototype to allow for comparisons with the previous versions.

¹<https://www.caixa.gov.br/Paginas/home-caixa.aspx>

For this updated prototype, our results show that users need less time, less interaction, higher satisfaction, and better effectiveness and user experience for most of the tasks we ask them to perform during our Usability Tests. For instance, to complete the task “Make a transfer to Banco do Brasil, Current Account, for an amount of R\$ 0.01”, users need 30 and 40 screen touches when using our updated prototype and the unmodified prototype of Caixa Tem, respectively. Besides that, users need 75 and 90 seconds to perform this task for both prototypes.

Therefore, we obtain a set of modifications resulting from Heuristic Evaluations and Usability Tests that provides a system with improved quality, ensuring better usability for users. Analyzing the results obtained from the assessed metrics, we conclude that each evolutionary version we develop for Caixa Tem showed an improvement in usability.

Our work is organized as follows: Section 2 presents the main concepts necessary for a better understanding of this work. In Section 4, we detail the methodology we used to conduct our research. In Section 5, we explain in detail the results obtained in the study. Section 3 discusses related work. Finally, in Section 7, we present our conclusion.

2 Background

In this section, we present the key concepts necessary for a better understanding of this work. In Section 2.1, we discuss digital public services in Brazil and their impact on the society. In Section 2.2, we explain what the Figma tool is, which is a fundamental element for our work. In Section 2.3, we elucidate the concept of high-fidelity prototypes. At last, in Section 2.4, we introduce the concepts of Heuristic Evaluation and Usability Testing.

2.1 Digital Public Services in Brazil

The technological advance in our daily lives is notably expressive. There are several conveniences and amenities that technology has been providing to society as a whole, with the increasing use of Information and Communication Technologies (ICTs) in our daily lives [Kosakowski, 1998]. Nowadays, in the private sector, companies that are not present in the digital realm are losing market share, emphasizing the growing influence of digital culture in society [Gere, 2009]. This trend brings a number of advantages, compelling the public sector to develop a management environment capable of keeping pace with these changes and also benefiting from the advantages generated by technology [Marchionini *et al.*, 2003]. In this context, there has been a gradual increase in the implementation of technologies in the Governo Eletrônico (e-gov) models adopted by Brazilian government [Motta, 2003].

For a brief overview of the last 20 years, we could highlight the launch of the Portal Governo Digital in 2000; the creation of the Portal da Transparência in 2004; the Portal da Inclusão Digital in 2006; the conduct of ICT surveys for the e-gov; the Acesso à Informação Law in 2011; the Marco Civil da Internet in 2014; the establishment of the Processo Nacional Eletrônico by Decree No. 8,539 in 2015; the Gov-

ernança Digital na Administração Federal with Decree No. 8,638 in 2016; the creation of the Sistema Nacional para a Transformação Digital by Decree No. 9,319 in 2018; and, more recently, the inauguration of the Gov.br Portal by Decree No. 9,756 in 2019. These milestones demonstrate the many changes in the Brazilian Public Administration, which has increasingly incorporated information and communication technologies into its operations [Cristóvam *et al.*, 2020].

The use of technology by the Brazilian government tends to provide several benefits for both the government and society. The proper use of Information and Communication Technologies (ICTs) has the potential to increase citizen participation, providing optimized access to public services, making such access less bureaucratic. Additionally, it can reduce government costs for carrying out these tasks and generate data and information that assist in decision-making and the creation of public policies for society [Jardim, 2000].

However, for the adoption and implementation of these types of technologies to be efficient, the government needs to ensure access for the entire population, minimizing differences, and guaranteeing the execution of social programs for the benefit of all [O’neill *et al.*, 2017]. For this disruptive digital advancement, the Public Sector needs to prioritize the most vulnerable segment of the Brazilian society [Viana, 2020]. Analyzing data from the Brazilian Institute of Geography and Statistics (IBGE), concerning trends in the use of these technologies, raises concerns as it may result in a portion of the population being excluded from access to public services and information. According to the IBGE, more than 25% of the Brazilian population does not use the internet in their homes. The main reasons for non-use of the internet, as per IBGE data, include lack of interest by 34% of the population, high service costs for 28.7% of the population, and unfamiliarity with internet usage for 22% of the population [IBGE, 2018].

These IBGE data took on a sad dimension amid the Coronavirus (Covid-19) pandemic when a considerable contingent of people in vulnerability faced difficulties accessing the emergency basic income benefit of six hundred Brazilian reais². This scenario happened due to issues with access and data filling in the application provided by the government in partnership with Caixa Econômica Federal (CEF)³, as well as problems related to the regularization of personal documents, such as the Cadastro de Pessoas Físicas - CPF [Cristóvam *et al.*, 2020]. Moreover, other Global South states such as India, also faced difficulties while adopting new technologies within the relation to their government [Gupta *et al.*, 2022]. Therefore, such issues are not exclusive to Brazil. These situations disproportionately affected those who were already more vulnerable and urgently needed the assistance and humanitarian action of the State [Viana, 2020].

In this context, we should seek strategies to foster digital democracy, aiming for the participation in a virtual public space that ensures access to public services for the entire population. This contributes to the elimination of “digital illiteracy”. To take this step, it is crucial to provide the entire population with access to the internet, technological devices,

²https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2020/lei/113982.htm

³<https://www.caixa.gov.br>



Figure 1. Image of the workspace of a project in Figma

and the personal capacity to use these technologies [Cunha and Miranda, 2013].

2.2 Figma

Figma is an online graphic editing platform that allows users to collaboratively develop vector projects and system prototypes. Widely used in both the market and academia, Figma provides functionalities for setting up interactions and navigation between pages designed within the platform. Primarily accessible via a web browser, it also has desktop and mobile versions, with the latter capable of mirroring prototype projects for interaction on mobile devices. This versatility makes Figma one of the leading UI/UX tools, providing a conducive environment for testing and enhancing graphic interface prototyping projects [Figma, 2011; Staiano, 2022].

In Figma, a project is developed through its graphic editor, which provides freedom for the user to create system screens as desired. Screens are fundamental elements in the prototyping process, and once created, the user can configure the navigation flow and usability using the prototyping menu. Figure 1 illustrates this environment, showing screens in the center, the top menu for adding elements such as screens and objects, the layers menu on the left, and the design and prototyping menus on the right, allowing adjustments to visual properties and configuration of interactions and navigation flows [Figma, 2011]. In this work, we use Figma to develop the prototypes that we evaluate and implement the improvements obtained through the study.

2.3 Prototyping

The prototyping process is essential in various projects, especially in software projects, where the need to represent something abstract is crucial [Budde *et al.*, 1990]. Prototyping in software projects is a fundamental practice to ensure the understanding of what is being developed, enabling validations and adjustments before implementation, making the process more efficient and cost-effective [Lichter *et al.*, 1994]. In software projects, prototypes are categorized into low, medium, and high fidelity [Rudd *et al.*, 1996]. Low-fidelity prototypes simplify the overall idea of the system, while medium-fidelity prototypes incorporate more details and design concepts. High-fidelity prototypes, on the other hand, resemble the final result, presenting all visual elements and allowing interactions, facilitating usability evaluation through testing [Rudd *et al.*, 1996].

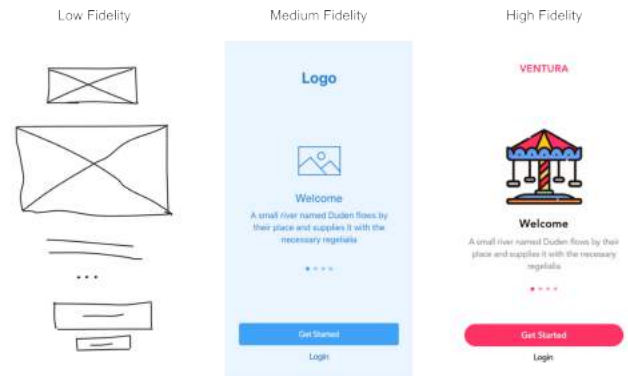


Figure 2. Example of prototype classification into low, medium, and high fidelity in a specific context

Figure 2 illustrates the representation of the three classifications. On the left, there is a simple sketch (low-fidelity prototype) made by hand, signaling the basic elements that compose the screen, such as text, images, and buttons. In the center (medium fidelity), we can see a better representation with some design concepts, using representations very close to the final elements, such as buttons, icons for images, and text. Finally, on the right, there is an example of a high-fidelity prototype, which already has the appearance of a finished system, with colors, logo, images, and full design customization of elements according to the visual identity of the system. In this study, we use high-fidelity representations.

2.4 Heuristic Evaluation and Usability Test

Heuristic Evaluation is a methodology developed by the researcher Jakob Nielsen [Nielsen, 1994]. He listed over 240 distinct problems that affect the usability of systems and developed a set of guidelines that, if met in a project, mitigate the possibility of these usability issues occurring [Nielsen, 1994]. As a result, he created a set of ten guidelines that have been widely used as a method for usability inspection in systems [Nielsen, 2005]. This method brings some advantages as it is quick and cost-effective, as it can be conducted by an inspection by one or more experts. The idea is to look for violations of the 10 heuristics proposed by Nielsen. Therefore, user participation is not required in this method [Nielsen, 1995].

Usability Testing, similar to Heuristic Evaluation, is an approach to assess the usability quality of a system or prototype [Lewis, 2006]. It involves allowing individuals representative of end users to interact with the system, navigating through the interface to perform selected typical tasks for the context, such as those considered critical or high-frequency. To ensure consistency in results, tests should be standardized, providing the same conditions to all users and instructing them to perform the same tasks. After the tests are completed, we can evaluate various criteria to analyze and compare results, identifying possible usability issues through the observation of difficulties faced by users [Riihiahho, 2018].

3 Related Work

In this section, we present the related work. The study by Karola Marky *et al.* [Marky *et al.*, 2020] addresses the usability of the Swiss voting scheme, identifying issues in the

adopted methodology and proposing improvements. They develop prototypes, conduct Heuristic Evaluations and Usability Tests to investigate usability, suggesting solutions for identified problems. After prototyping the solutions, they conduct new tests, evaluating performance in specific metrics. Our work follows a similar methodology to Karola Marky's, although it differs in intrinsic aspects related to online voting systems. In the context of Caixa Tem, we evaluate specific issues, such as the impracticality of applying methodologies that involve modifying data to ensure the confidentiality of participants' votes in the evaluations.

Cigdem Altin [Gumussoy, 2016] conducts a study with the aim of establishing usability guidelines for banking systems through heuristic analyses. The research examines three banking software projects, identifying usability issues through Heuristic Evaluations and categorizing them by severity. The issues are analyzed and structured to create guidelines that address the main issues identified in the study. Although this work provides valuable contributions, it differs from ours by exclusively employing Heuristic Evaluation, not allowing for the evaluation of the system by end users, which limits the identification of specific problems addressed in methods such as Usability Testing.

Mutlaq B. Alotaibi's study [Alotaibi, 2016] aims to list and compare the usability of various mobile applications in Saudi Arabia, distinguishing between market mobile systems (M-business) and governmental ones (M-government). Thirty-six applications, equally divided between M-business and M-government, were analyzed to represent the main mobile options in the country. Usability assessment was conducted through a questionnaire based on Nielsen's heuristics, adapted into a checklist with eleven items. Thirty-six participants evaluated each application according to the criteria defined in the questionnaire. The author concludes that M-business applications exhibit better usability compared to M-government ones. Unlike our work, this study employs a specific technique, adapting Heuristic Evaluation into a questionnaire combined with in-app interaction, similar to a simple Usability Test.

The work conducted by Layla Hasan [Hasan, 2013] investigates the usability of three websites from public universities in Jordan. The study employs Heuristic Evaluations to identify a total of 34 categories of usability issues occurring in these systems. Ultimately, she presents a list detailing these 34 categories, along with exemplifying these problems on the websites. This list serves as a guide for addressing these issues or preventing them in other projects. The research is limited to identifying and categorizing the main problems found on university websites, unlike our work, which goes further by implementing and testing to assess the consequences of addressing these issues.

The work carried out by Janet Chisman [Chisman *et al.*, 1999] aims to investigate the usability of the libraries at Washington State University (WSU). To achieve this, a set of Usability Tests was conducted to identify usability issues within the library system. After the tests, a series of recommendations were developed to address the identified issues in the study, which were implemented in a subsequent version of the university library system. This study, unlike ours, is also limited to identifying problems and proposing cor-

rections, but in this case, using Usability Tests rather than Heuristic Evaluation.

4 Research Method

In this section, we present the methodology we apply in this study. In Section 4.1, we outline the entire flowchart of how we conducted the research. In Section 4.2, we explain the reason for selecting the Caixa Tem app, providing a general overview of the app and outlining the functionalities considered in the study. In Section 4.3, we present the hypotheses studied in the work and the metrics used to evaluate these hypotheses. Additionally, in Section 4.4, we present the stages of evaluation and prototyping, detailing the steps followed in the study. At last, we provide the documents used as inputs for the construction of the research in our online Appendix [Vieira and Andrade, 2024].

4.1 Research Procedure

To detail the research procedure adopted in this study, we present the flowchart in Figure 3. It outlines the steps we take during our research.

1. We initiated by selecting the system for our research. To commence this process, we conducted an analysis to decide which public system would be the focus of our study. Thus, we selected the Caixa Tem app [Caixa Econômica Federal - CEF, 2020]. In Section 4.2, we delve more into the reasons for this choice.
2. We formulated hypotheses that we could evaluate during the research, serving as a means to measure the results obtained in this study (Section 4.3).
3. We selected metrics to assist in validating or refuting the previously considered hypotheses. Section 4.3 provides details about the hypotheses and metrics addressed in this study.
4. We prototyped the original system using Figma (Section 2.2), aiming for the prototype to be as faithful as possible to the original Caixa Tem app. Thus, users interacting with the Original prototype should have a similar experience to using the real app. In Section 4.4.1, we detail how we develop the entire prototyping.
5. We initiated the usability evaluation steps of the Caixa Tem app. Initially, we conducted a Heuristic Evaluation, aiming to identify possible usability problems and proposed solutions for those. We explain in detail how we conduct the Heuristic Evaluation in Section 4.4.2.
6. We considered the results of the Heuristic Evaluations from the previous step to make modifications to the Original prototype, seeking to resolve or mitigate usability issues we find. Thus, we aimed to improve the quality of the system's usability. Therefore, at the end of this step, we had the Redesign prototype with the improvements learned in the Heuristic Evaluation.
7. We conducted Usability Tests, where users interact with both the Original and the Redesign prototypes. This allowed us to extract data on the selected metrics. Thus, we evaluated the usability of both prototypes. In Sec-

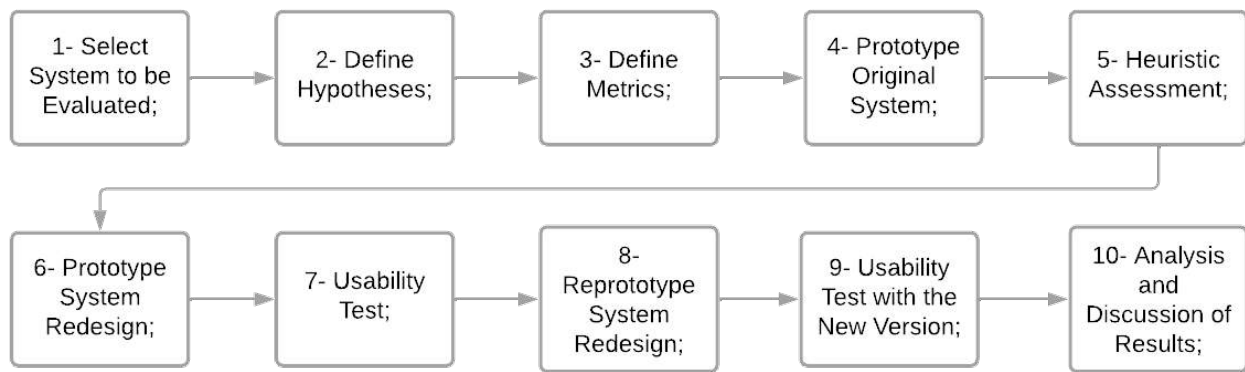


Figure 3. Flowchart of the steps of the research procedure adopted for the study

tion 4.4.3, we provide more details about the Usability Tests.

8. We conducted a process similar to the sixth step. However, we considered the results of the Usability Tests, not the Heuristic Evaluations, to make modifications to the Redesign prototype. The aim is to address or mitigate usability issues we found, seeking to further improve the quality of the system's usability. At the end of this step, we implemented a new version of the prototype, which we call the Final Design.
9. We conducted another round of Usability Tests. Differently from the eighth step, users interacted only with the Final Design prototype. In Section 4.4.3, we provide more details about the usability tests.
10. We discussed and analyzed the results obtained from the usability evaluations we conducted in the previous steps. Sections 5 and 6 provide more information.

4.2 Caixa Tem Application

In order to select a Brazilian public system, we analyze *Urna Eletrônica*⁴, *Meu Imposto de Renda*⁵, *SOUGOV*⁶, *SigProj*⁷, and *SigPAC*⁸. However, we ultimately choose Caixa Tem due to the visibility the system had at the time of the COVID-19 pandemic.

The Caixa Tem mobile app was launched in April 2020 as the primary means for the population to access the Brazilian Emergency Aid during the COVID-19 pandemic [Cardoso, 2020]. The critical need for access to the app, especially by the more vulnerable segment of society, emphasized the importance of usability to ensure the full utilization of services by the Brazilian public [Cristóvam *et al.*, 2020]. However, existing data indicates that the majority of those in need did not have internet access to register and obtain the aid [Viana, 2020]. This research also highlights that a portion of the Brazilian population is illiterate or has some form of disability [IBGE, 2018, 2019]. Besides the use during the COVID-

19 pandemic, Caixa Tem is still widely used nowadays for a number of different purposes⁹.

In this context, our study focuses on the usability of the Android version of Caixa Tem app, analyzing potential issues and proposing solutions [Cunha and Miranda, 2013]. Caixa Tem provides basic banking functionalities, as well as features such as payment at the lottery, cardless withdrawal, access to NIS¹⁰ and Bolsa Família¹¹, presenting them in a unique way by simulating a conversation history in a chat format [Cunha and Miranda, 2013]. The goal is to enhance the quality and usability of the system [Cunha and Miranda, 2013].

This approach used by the app aims to enhance the usability of the system, making it similar to popular messaging apps such as WhatsApp and Telegram. In Section 5, we discuss whether this approach indeed enhances the usability of the app in the users' opinion.

For this research, we consider typical features of the Caixa Tem. Thus, we select ten features: *Balance Inquiry*, *Payment at the Lottery*, *Cardless Withdrawal*, *Bolsa Família*, *Phone Recharge*, *Pix*¹², *Transfer*, *Bill Payment*, *Payment at the Card Machine*, and *Virtual Debit Card*.

4.3 Hypotheses and Metrics

In this section, we present the hypotheses we test in Section 4.3.1 and the metrics we select to test those hypotheses in Section 4.3.2. Finally, in Section 4.3.3, we present a summary table of hypotheses and metrics.

4.3.1 Hypotheses

For this study, we formulate five hypotheses, as follows:

- **H1** - The Final Design prototype provides greater user satisfaction compared to the Original prototype;
- **H2** - The Final Design prototype requires less time on average for the user to perform tasks compared to the Original prototype;

⁴<https://www.justicaeleitoral.jus.br/urna-eletronica/>

⁵<https://www.gov.br/receitafederal/pt-br/assuntos/meu-imposto-de-renda>

⁶<https://sougov.sigepe.gov.br/sougov/>

⁷<http://sigproj.ufrj.br/>

⁸<https://www.gov.br/transportes/pt-br/centrais-de-conteudo/sigpac-png/view>

⁹<https://www.caixa.gov.br/caixatem/perguntas-frequentes/Paginas/default.aspx>

¹⁰<https://www.gov.br/pt-br/servicos/consultar-dados-do-cadastro-unico-cadunico>

¹¹<https://www.gov.br/mds/pt-br/acoes-e-programas/bolsa-familia>

¹²<https://www.bcb.gov.br/estabilidadefinanceira/pix>



Figure 4. Screenshots of some of the main screens of the Caixa Tem app

- **H3** - The Final Design prototype requires, on average, a lower number of user interactions to perform a task compared to the Original prototype;
- **H4** - The Final Design prototype provides a better user experience compared to the Original prototype;
- **H5** - The Final Design prototype ensures better effectiveness in user-performed tasks compared to the Original prototype;

Hypothesis **H1** addresses user satisfaction when interacting with the prototypes. Therefore, testing **H1** allows evaluating which prototype provides greater satisfaction to users. Hypothesis **H2** considers the time spent by the user to perform tasks in the application. Therefore, testing **H2** allows evaluating which prototype takes less time to complete tasks. Hypothesis **H3** addresses the number of interactions the user needs to perform with the system to complete a task. Therefore, testing **H3** allows evaluating which prototype requires a lower number of interactions. Hypothesis **H4** addresses the user experience when interacting with the prototypes. Therefore, testing **H4** allows evaluating which prototype provides a better experience for users. Finally, Hypothesis **H5** considers the effectiveness of the prototypes used by users. Therefore, testing **H5** allows evaluating whether users can successfully complete tasks in the system or not.

4.3.2 Metrics

After defining the hypotheses considered in this study, we define the metrics we use to validate or refute the hypotheses presented in Section 4.3.1. We calculate the metrics presented in this section with the data obtained from the Usability Tests. In this way, we select a set of five metrics to assess the usability of the studied prototypes: *time efficiency*, *interaction efficiency*, *satisfaction*, *user experience*, and *effectiveness*. We describe each metric below.

Time Efficiency. This metric consists of measuring the time to perform a task related to each of the selected functionalities [Marky *et al.*, 2020]. During the tests, participants are encouraged to verbally signal when they are starting a new task and when they consider that they have completed that specific Task. Thus, we could measure the time spent on the particular task [Boren and Ramey, 2000].

Interaction Efficiency. It involves analyzing the number of user interactions with the system to complete each of the tasks in the Usability Test. Since this is a study with a mobile application, we assess the number of times the user taps on the smartphone screen. We measure this metric through screen recording during the test, capturing the 'clicks' the

user makes, thus quantifying the number of interactions required for each task.

Satisfaction. It consists assessing the user's satisfaction with the usability of the evaluated system or prototype, making it a subjective criterion. In this study, we measure the subjective usability satisfaction through the SUS (System Usability Scale) [Brooke *et al.*, 1996], which consists of a form filled out by users after completing the usability test. Once participants complete the form, it results in a numerical value ranging from 0 to 100 points. Higher scores indicate better subjective usability, allowing us to measure and evaluate user satisfaction with the evaluated prototype.

User Experience. It regards assessing how the user's experience was while using the system or prototype. Similar to the satisfaction metric, the measurement of the experience also considers subjective criteria. In this study, we measure the user experience through the UEQ (User Experience Questionnaire), analyzing six scales: 1) attractiveness, 2) transparency, 3) efficiency, 4) control, 5) stimulation, and 6) novelty. The questionnaire is filled out by users after interacting with the prototypes in the usability test. Once completed, a graph is generated comparing the results of the studied prototypes on the six scales described above, allowing for the evaluation of the user experience in relation to the evaluated prototypes [Laugwitz *et al.*, 2008].

Effectiveness. It considers whether the user was able to successfully complete each of the tasks requested in the usability test or not. This metric is measured from the test recordings, where the researcher assesses from the recording whether the user correctly completed each task or not, thus signaling the data for this metric. This allows measuring the user's effectiveness in performing each task, as well as analyzing the factors that influence low effectiveness, so that they can be corrected in the future [Marky *et al.*, 2020].

4.3.3 Relationship between Hypotheses and Metrics

Table 1 synthesizes the entire scheme needed to test the hypotheses we raise in this study. For example, to test hypothesis **H1**, it is sufficient to use the Satisfaction metric (Section 4.3.2), which is measured through the SUS scale. Therefore, we assess the average SUS scores in the usability tests for each prototype, to validate or refute the hypothesis in question. For the other hypotheses, we follow the same evaluation pattern based on Table 1.

Hypotheses	Metric	Data collection	Data analysis
H1	Satisfaction	Measured through the application of the SUS questionnaire to users after conducting Usability Tests.	To evaluate the average score generated by the SUS questionnaire for each of the prototypes assessed.
H2	Time Efficiency	Measured through the recording of Usability Tests, allowing for the assessment of the time spent on each task.	To analyze the average time to complete all test tasks, as well as the time spent on each individual task.
H3	Interaction Efficiency	Measured through the recording of Usability Tests, which signals the “clicks” the user made on the mobile device screen.	To evaluate the average number of interactions required to perform the test tasks, as well as the number of interactions for each task.
H4	User Experience	Evaluates the scores obtained from the User Experience Questionnaire (UEQ), filled out by users after completing Usability Tests.	To analyze the comparative graph between the prototypes, which compares each of the 6 criteria assessed by the UEQ.
H5	Effectiveness	Evaluates the quantity of tasks successfully completed by users in Usability Tests.	To analyze the percentage of successfully completed tasks in each of the prototypes.

Table 1. Summary of hypotheses, with their respective metrics, and the way to measure and evaluate each one



Figure 5. Home screen of the three prototype versions

4.4 Exploratory Study

To evaluate the Caixa Tem usability, we use high-fidelity prototypes, as presented in Section 2.3. Therefore, we prototype three versions of the Caixa Tem application: an original version identical to the real application and two versions that implement improvements learned from usability evaluation. We call these prototype versions Original, Redesign, and Final Design. Section 4.4.1 provides detailed information on how we conducted the prototyping. To assess the usability of the prototypes, we employed two techniques. First, we conduct Heuristic Evaluation in one round, where experts examined the Original prototype (Section 4.4.2). Finally, we conduct two rounds of Usability Tests with users to evaluate the Redesign and Final Design prototype versions (Section 4.4.3).

4.4.1 Prototyping

In total, we prototype three distinct versions of the Caixa Tem application: Original, Redesign, and Final Design. The Original prototype is faithful to the original application, serving as the control version in our study. Its purpose is to replicate the usability provided by the real application as closely as possible, ensuring that users interacting with this version have an experience very similar to using the actual Caixa Tem application. The Redesign prototype is based on the Original but incorporates the improvements identified during the Heuristic Evaluations. The Final Design prototype, on the other hand, builds upon the Redesign version and incorporates the improvements identified in the first round of Usability Testing.

We use the Figma tool [Figma, 2011] to prototype our three design versions. Indeed, we prototype screens and user interactions, replicating as closely as possible the user experience of the actual Caixa Tem application.

In Figure 5, we observe some of the progress that the pro-

toype underwent during the study. From the Original version to the Redesign, we see mainly the standardization in iconography that we develop through improvements identified during the Heuristic Evaluation. In the transition from the Redesign to the Final Design, we notice the change in the text of the balance display functionality button, which we identified as an issue during Usability Testing. Thus, we apply it as a correction to improve the effectiveness of the related task.

4.4.2 Heuristic Evaluation

In this work, we use Heuristic Evaluation [Nielsen, 2005] to quickly and cost-effectively identifying usability issues and suggesting corrections for these identified problems. In this context, it is important that experts run the Heuristic Evaluation in order to obtain a useful result [Nielsen, 1995]. For this study, we select students from the Bachelor of Computer Science program at (omitted for anonymity), who have completed the Human-Computer Interaction course.

In this way, we recruit 16 Undergrad Computer Science students. For each student who willingly contributed to the research, we provide a guide with supporting content and instructions on how they should conduct this evaluation. This material includes a video where we briefly present the purpose of the research and review the basic concepts of Heuristic Evaluation, showcasing Nielsen’s ten heuristics [Nielsen, 2005]. At the end of the video, we provide instructions on how participants should document and submit the data from their evaluation.

For this purpose, along with the video, we provide a document to be used for documenting the evaluation. For each violation of one of the ten heuristics, the evaluator was instructed to indicate which heuristic was violated, explain how the violation occurred, assess the severity of this violation in the usability of the system, provide the necessary correction recommendation, and attach images exemplifying the violation. The evaluation was to be conducted on a mobile device using the Original Prototype. Each evaluator carries out the Heuristic Evaluation individually.

After the students complete the Heuristic Evaluation, they send us their results so that we conduct an analysis of all the identified violations. The goal of this step is to filter the problems that are indeed useful for this study. On the other hand, we discard those identified issues that are not useful (e.g., not related to usability). Additionally, prioritize violations that

have the potential to improve the usability of the Caixa Tem application, which are the main focuses of our study, and that are also feasible to be implemented and tested in our prototypes.

4.4.3 Usability Test

For the efficient execution of Usability Tests, it is crucial to pay attention to two points: 1) all participants must have the same conditions for conducting the test so that this does not influence the obtained results, and 2) the users selected to perform the tests should be representative users, meaning users whose profiles align with the target users of the application [Riihiho, 2018]. In the study, we conduct two rounds of Usability Tests. In the first round, we test the Original and Redesign prototypes. In the second round, we test the Final Design prototype.

In this context, we devise a standardization for the tests, aiming to avoid bias on the final result. We run the tests in-person and individually, with only the participant and the first author present. We use the same notebook and smartphone in all tests. The notebook containing the instructions and the smartphone running the prototype. We record the smartphone screen and audio during the interaction to facilitate data collection. During this phase, we encourage participants to think aloud to better understand their interaction with the interface [Boren and Ramey, 2000]. The first author observes the participant's behavior and expressions, documenting comments when necessary to assess the demonstrated experience.

All tests follow a standard script, conducted as follows:

- **1) Welcome and Demographic Data Collection** - We begin by thanking the participant for their voluntary collaboration in the research, providing a brief overview of the study's purpose without revealing all the details to avoid influencing the test. We ensure clarifications at the end of the process. We explain the test, providing the script and an estimated duration. We present the consent form, emphasizing voluntary participation, authorization for screen and audio recording, and the anonymous use of test data. We conclude with a brief interview to gather demographic information.
- **2) Instruction and Interaction with the Prototype** - The first author guides participants to interact with the prototype, encouraging verbal expression during the interaction. Participants are prompted to try to complete the tasks on their own, with the option to seek assistance from the first author if encountering difficulties. They are asked to verbally communicate the start and completion of each task. Participants are guided to consider the real need to complete the tasks for personal purposes. After the instructions, we present a list of ten tasks to be performed in the system:

1. Check the balance in your account;
2. Use the virtual debit card and generate your security code;
3. Make a bill payment;
4. Generate a code for payment at the lottery;
5. Generate a QR code to receive a Pix;

6. Make a payment via Pix using the phone Pix key;
 7. View the payment schedule for Bolsa Família;
 8. Generate a code for cardless cash withdrawal;
 9. Recharge your phone with TIM, for an amount of R\$ 20.00;
 10. Make a transfer to Banco do Brasil, Current Account, for an amount of R\$ 0.01.
- **3) Form Completion** - After interacting with the prototype, participants are guided on how to respond to two usability questionnaires, the UEQ and SUS. We explain the purpose of each questionnaire and provide considerations for proper completion: 1) express your opinion fairly, 2) respond spontaneously without overthinking, 3) mark an answer even in case of uncertainty, 4) there are no right or wrong answers, and 5) evaluate only based on the experience with the prototype. Participants fill out the form digitally on the notebook used in the test.
 - **4) Final Questionnaire** - After participants complete the form, the researcher conducts some quick oral questions. The focus is to assess whether the participant would like to use the system they interacted with in the test frequently, and to investigate the reasons why or why not. We also inquire about the participant's opinion on the application's philosophy of simulating a chat, asking whether, in their opinion, this philosophy aids in usability. Finally, we allow the participant to freely express any criticism, comment, suggestion, or opinion about the tested prototype.
 - **5) Ending** - The first author thanks the participant once again for their collaboration, provides a comprehensive explanation of the entire study being conducted, and offers to address any questions the participant may have.

Participants for usability tests are selected based on pre-established criteria to ensure the inclusion of representative users in the study. In this case, we defined three potential user profiles with a higher likelihood of using the Caixa Tem application: 1) Individuals under the age of 30, who used it for personal needs or to assist a family member or close person having difficulty with digital systems. 2) Individuals aged between 30 and 50, representing the adult audience requiring the use of features provided by the Caixa Tem application. 3) Individuals aged 50 and above, representing an older audience requiring the use of features provided by the Caixa Tem application. It is important to note that for each usability evaluation stage, whether through Heuristic Evaluation or Usability Testing, distinct participants are selected, and the same person is not repeated in different stages.

Therefore, to conduct the first round of usability tests, where we evaluate the Original prototype and the Redesign prototype, we select 12 participants. There are four participants in each of the three age profiles described above. These 12 participants were randomly drawn to form two groups, each with six participants, with each group having two participants from each of the three age profiles mentioned above. Each of the two groups interacts with one of the prototypes under analysis. Table 2 summarizes this scenario.

As for the second round of usability tests, we evaluated the third version of the prototype, the Final Design prototype.

Participants above 50 years old		Participants between 30 e 50 years old		Participants below 30 years old	
Participant	Prototype version	Participant	Prototype version	Participant	Prototype version
1	Redesign	5	Original	9	Original
2	Original	6	Redesign	10	Redesign
3	Redesign	7	Redesign	11	Redesign
4	Original	8	Original	12	Original

Table 2. Presentation of the division of participants by the age profile that tested each prototype

In this case, we selected six participants, ensuring that there were two participants from each of the age profiles we presented (Profile 1 < 30 years; Profile 2 between 30 and 50 years; Profile 3 > 50 years). Thus, there were two participants under 30 years, two between 30 and 50, and two above 50 years. This additional round was necessary because, after identifying usability issues in the first round of Usability Tests, even after conducting Heuristic Evaluations, we proposed improvements for these identified issues. This test serves precisely to evaluate the resulting changes in this final version of the prototype.

5 Evaluation

In this section, we present the results obtained from the Heuristic Evaluations (Section 5.1). Additionally, in Section 5.2, we present the results obtained in the two rounds of Usability Tests. At last, we discuss the threats to validity regarding our work in Section 5.3.

5.1 Heuristic Evaluation

In total, each of the 16 participants performs a Heuristic Evaluation. This results in a total of 49 violations of Nielsen's Heuristics [Nielsen, 2005]. In this context, we filter these violations to check whether they are useful for this study. Out of these 49 violations, we consider 25 for this study. We discard the remaining 24, as we detail next.

In Table 3, we summarize the violations, detailing their quantity, those considered and those discarded for each heuristic. Thus, we list Nielsen's 10 Heuristics and indicate how many violations of each Heuristic we discard and how many we consider for this study. For example, when analyzing the first Heuristic described as *H1 - Visibility of System Status*, we see that we consider three violations of this heuristic, while we discard four of them.

5.1.1 Severity of Considered Violations

The evaluator identifies violations by indicating the violated Heuristic, describing the flaw, proposing solutions, and assessing severity based on frequency, impact, and persistence. Severity is classified into four levels: 1) Cosmetic issue, low impact, and optional correction; 2) Minor issue, with minimal impact and low priority; 3) Major issue, with considerable impact and high priority; 4) Catastrophic issue, with extreme impact and maximum priority, preventing product release.

Thus, in Table 4, we present the severity of issues in the 25 violations that we consider in this study. We observe that

most violations are classified by evaluators as minor or major issues, quantifying 10 and 14 violations, respectively. For the other categories, there is only one violation classified as cosmetic, and none as catastrophic.

5.1.2 Discarded Violations

We justify and classify the 24 discarded violations into four groups, representing the category of motivation for their dismissal:

- *1) External Factors:* representing usability issues that fall outside the scope of the application. An example of a violation indicated and classified in this category is related to the smartphone keyboard that appears in the prototype, being a configuration of the device's operating system and not of the application itself;
- *2) Repeated Violations:* as the evaluations were conducted in parallel, more than one evaluator has the possibility of encountering the same violation. For these cases, we considered all proposed improvement suggestions, implementing the one we assessed to demonstrate higher quality. Thus, we classified only one of the repeated violations as a considered violation, discarding the others and placing them in this category;
- *3) Imprecise or Mistaken Violations:* some violations are considered imprecise or mistaken because, in some cases, they do not contextualize and clarify the addressed problem. In other cases, it may not actually be a violation, and in some, the proposed improvement suggestion might even worsen usability instead of improving it. For example, a case was the suggestion to group several functionalities under a single menu option, which could make it more difficult for users to find functionalities;
- *4) Figma Limitations:* some violations cannot be feasibly implemented in Figma, which makes it impossible to test the effect of these corrections; therefore, they were also discarded. An example of a violation in this category is having the last-used functionality always appear at the top of the list in the main menu. Implementing this in Figma becomes impossible as it would require prototyping all possible combinations of function ordering.

In Table 5, we present the quantity of dismissals by category. We categorize three dismissals as "External Factors," related to the appearance of the smartphone keyboard in the prototype which is configured in the device's operating system not in the Caixa Tem application. Additionally, we dismiss five violations due to repetition, i.e., more than one

Heuristic	Number of Considered Violations	Number of Discarded Violations
H1 - Visibility of system status	3	4
H2 - Match between system and the real world	1	0
H3 - User control and freedom	4	0
H4 - Consistency and standards	4	2
H5 - Recognition rather than recall	4	1
H6 - Flexibility and efficiency of use	1	4
H7 - Aesthetic and minimalist design	1	7
H8 - Error prevention	2	3
H9 - Help users recognize, diagnose, and recover from errors	5	3
H10 - Help and documentation	1	0

Table 3. Detailed presentation of the quantity of violations considered and discarded for each Nielsen Heuristic

Severity	Number of Violations
Cosmetic issue	1
Minor issue	10
Major issue	14
Catastrophic issue	0

Table 4. Presentation of the severity of the 25 violations considered in the study

Category	Number of Violations
External Factors	3
Repeated Violations	5
Imprecise or Mistaken Violations	12
Figma Limitations	4

Table 5. Presentation of the quantity of discarded violations in each of the four classes of discarded violations in the study

evaluator identified the same violation. The majority of dismissals, totaling 14 violations, occurred due to imprecise or misguided justifications. An example included in this category is the suggestion to modify screens of specific functionalities, considered objective and aligned with the philosophy of chat simulation. Finally, we discard four violations due to limitations in Figma. An example in this category is the suggestion to rearrange the main menu based on the last used functionality, considered costly to implement in Figma due to the various possible ordering combinations.

5.1.3 Example of Violation

Figure 6 illustrates an example of a violation identified by one of the evaluators during the Heuristic Evaluation. In it, we can analyze a table where the evaluator fills in all the data about the evaluation, followed by a screenshot of an attached screen, illustrating the occurrence of the identified violation. Initially, the evaluator indicates which of Nielsen’s ten Heuristics the problem in question violates, followed by a brief description detailing the entire violation. The severity of this violation is also indicated according to the evaluator. Finally, a modification suggestion is provided to correct or mitigate this violation. All other violations are documented following the same structure presented in this example.

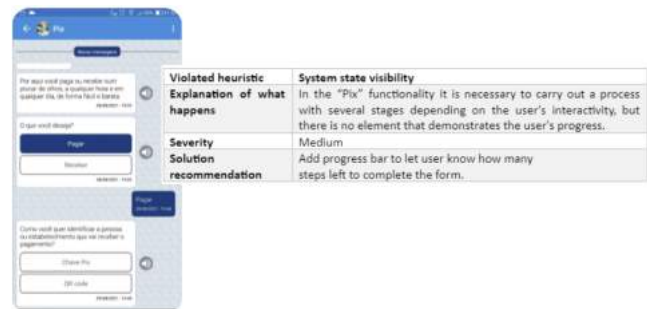


Figure 6. Example of a violation identified in the Heuristic Evaluation by an evaluator

5.2 Usability Test

In this section, we present the results of the two rounds of Usability Tests we conduct on Caixa Tem prototypes. In the first round, we carry out the Usability Tests on the Original and Redesign prototypes. We detail the results of this initial round in Section 5.2.1. Moving on to the second round, we conduct Usability Tests on the Final Design prototype. The results of this second round are described in Section 5.2.2.

5.2.1 First Round

In this section, we present the results for the first round of Usability Testing, as well as the profile of the participants who underwent this initial round. Finally, we highlight the key insights and changes gleaned from this Usability Testing round.

Participants. To conduct the first round of usability tests, we recruit 12 participants through direct contact. In this context, we segment them into three groups based on age range. The first group consists of individuals aged 18 to 29, the second group includes those aged 30 to 49, and the third group comprises participants above 50 years old (as described in Section 4.4.3). We select the 12 participants to ensure there are four participants in each of the three age groups defined above. Within each group, we randomly assign the four participants to determine which prototype each would test. Thus, in each group, two participants test the Original prototype, and two test the Redesign prototype. Their average age is 37.25 years (*Standard Deviation* \approx 14, *Median* = 31, *Min* = 21, *Max* = 57). Of the participants, 33.3% (N=4) identified as female, while 66.7% (N=8) identified as male. Furthermore, 33.3% (N=4) completed high school, 41.7% (N=5)

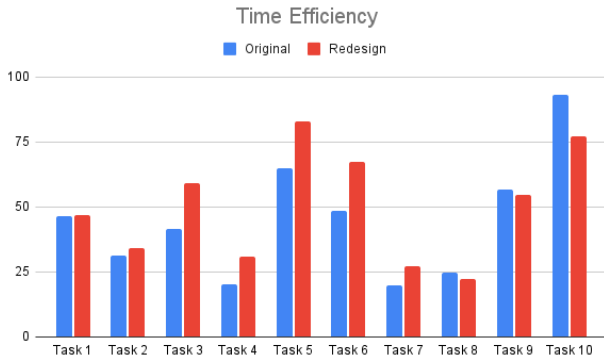


Figure 7. Time Efficiency First Round Graph

completed higher education, and 25% (N=3) completed post-graduate studies. All participants are employed. All selected participants had prior experience with the Caixa Tem application, with 83.3% (N=10) using the application solely for emergency aid, 8.3% (N=1) using it for both emergency aid and PIS, and 8.3% (N=1) using it for both emergency aid and labor-related finances.

Time Efficiency. Participants using the original prototype require an average of 44.71 seconds for each task (*Standard Deviation* \approx 22.93 seconds, *Median* \approx 43.91 seconds, *Min* = 19.83 seconds, *Max* = 93.33 seconds). On the other hand, participants using the Redesign prototype need an average of 50.26 seconds for each task (*Standard Deviation* \approx 21.38 seconds, *Median* \approx 50.66 seconds, *Min* = 22.5 seconds, *Max* = 82.83 seconds). Thus, the Original prototype shows a 12.41% greater time efficiency compared to the Redesign prototype.

The graph in Figure 7 illustrates the time efficiency of each prototype for each of the 10 proposed tasks. Thus, we observe that, on average, the Redesign prototype requires a slightly longer time compared to the Original. However, this difference varies from Task to Task. For instance, when examining Task 1, which involves checking the account balance, both prototypes require a very similar amount of time, as it is a straightforward task with a clear flow. On the other hand, when analyzing Task 5, which involves generating a QR Code to receive a Pix payment, a more significant difference in the time needed for each prototype is evident. This is due to the fact that it is one of the functionalities that underwent modification, indicating an increased time requirement to complete the task.

Interaction Efficiency. Participants using the Original prototype require an average of 15.41 interactions for each task (*Standard Deviation* \approx 9.70, *Median* \approx 11.58, *Min* = 6, *Max* = 39.5), while participants using the Redesign prototype need an average of 16.35 interactions for each task (*Standard Deviation* \approx 8.99, *Median* \approx 14.33, *Min* = 6.16, *Max* = 32.33). Thus, the Original prototype shows an interaction efficiency 5.70% higher compared to the Redesign prototype.

The graph in Figure 8 shows the interaction efficiency of each prototype for each of the 10 proposed tasks. Thus, we observe that, on average the Redesign prototype require slightly more interactions with the system compared to the Original. However, this difference varies from task to task. For example, when examining Task 8, which involves generating a code for cardless withdrawal, both prototypes require

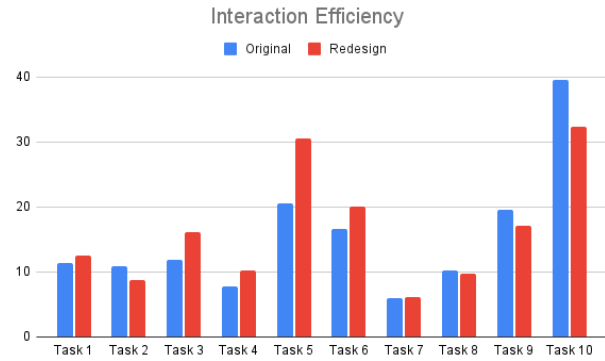


Figure 8. Interaction Efficiency First Round Graph

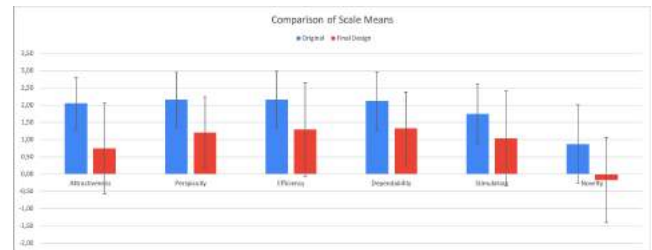


Figure 9. User Experience First Round Graph

a very similar number of interactions, as it is a simple task with a similar flow in both prototypes. On the other hand, when analyzing Task 5 again, which involves generating a QR Code to receive a Pix payment, a more significant difference in the number of interactions needed for each prototype is evident. This is due to the fact that it is one of the functionalities that underwent modification, as mentioned earlier, demonstrating an increased need for system interaction to complete the task.

Satisfaction. Users testing the Original prototype rated the SUS scale with an average of 77.08 points (*Standard Deviation* \approx 14, *Median* \approx 77.5, *Min* = 60, *Max* = 95), while users testing the Redesign prototype rated the SUS scale with an average of 73.75 points (*Standard Deviation* \approx 12.32, *Median* \approx 77.5, *Min* = 57.5, *Max* = 87.5). Thus, the Original prototype has a slightly higher score compared to the Redesign prototype.

User Experience. Figure 9 illustrates the results of the UEQ. Analyzing each scale through the t-test, which is the standard statistical method adopted by the UEQ method to assess the significance of differences between the analyzed samples, represented in the graph as the vertical black bar, we conclude that there are no significant differences in any of the six scales. This happens because, despite observing variation in the scales on the graph, we can only assert a significant variation when there is no intersection between the black bars representing the t-test in each analyzed scale. Thus, upon analyzing the graph, we observe that in all scales, the bars representing the t-test intersect. Therefore, we assert that both prototypes provide an equivalent user experience.

Effectiveness. Since each participant has ten tasks to perform, we calculate effectiveness based on the percentage of successful completion of the proposed tasks. We also separately analyze the effectiveness of each task, comparing the two prototypes under study. Looking at the overall results, the Original prototype shows an effectiveness rate of 81.66%.

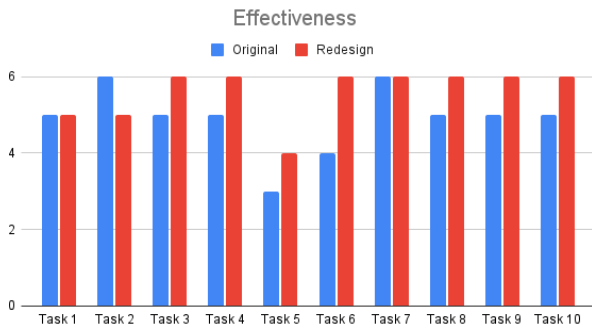


Figure 10. Effectiveness First Round Graph

More specifically, participants successfully complete 49 out of the 60 tasks. In contrast, the Redesign prototype demonstrates an effectiveness rate of 93.33%. Additionally, participants successfully complete 56 out of the 60 tasks. Thus, the Redesign prototype exhibits an effectiveness 14.28% higher than the Original prototype.

The graph in Figure 10 illustrates the effectiveness of each prototype for each of the 10 proposed tasks. Thus, we observe that despite the Redesign prototype showing higher overall effectiveness on average, there is variation when analyzing tasks separately. For example, in Tasks 1 and 7, which involve checking the account balance and consulting the payment calendar for Bolsa Família, both prototypes exhibit identical effectiveness. This occurs because these functionalities are exactly the same in both prototypes, with no modifications to their flows. On the other hand, when analyzing Task 6, which involves making a payment via Pix, we see that the Redesign prototype shows higher effectiveness. This is due to the fact that some users find certain options confusing in the flow of this task in the Original application. We address this issue for the Redesign prototype.

At the end of this first round of Usability Testing, we identify some corrections that we could implement in a new prototype version to enhance usability. In summary, in three out of the ten tasks, there are points for potential usability improvements. Below, we list the functionalities and the modifications:

- **Balance Inquiry:** Some participants, especially those in the group aged over 50, encountered difficulty in locating the balance. This functionality, unlike others, is situated in the header of the main menu under the option “Show Balance.” Therefore, we modify the button text to “Show Account Balance” to make the button’s function more apparent. We can observe the change in the Figure 11a;
- **Pay Your Bills:** There was an error that we did not notice during the prototyping process. In the header of the screen, it displays “Transfer Money” instead of “Pay Your Bills”, which could cause uncertainty for the user. Therefore, we correct this error for the new prototype version. Figure 11b shows this modification;
- **Receive Pix:** The functionality regards generating a QR Code to receive a Pix payment. Upon analyzing the test recordings, we observe that the majority of users face difficulties. Many users expressed frustration with this process. In the Original prototype, the screen for generating the QR Code differs from others, as shown in

Figure 12. The simulation of a chat, used for interactions in general, confused users, leading them to click in inappropriate places. In response, we modify the flow of this task, making it consistent with the chat pattern adopted in other functionalities of Caixa Tem.

5.2.2 Second Round

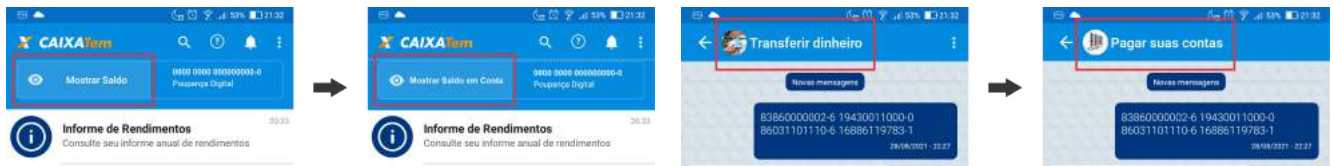
In this section, we present the results for the second round of Usability Testing, as well as the profile of the participants.

In this round, we conduct Usability Testing solely on the third version of the prototype (Final Design). This approach allows us to compare the improvements we implement in this version with the other two versions already tested: Original and Redesign prototypes. However, the second round follows the same structure as the first.

Participants. To conduct the second round of usability tests, we recruit six participants through direct contact, maintaining the same methodology applied in the first round of tests. Thus, we segment the participants into three groups based on age range: the first group comprises individuals aged 18 to 29, the second group includes those aged 30 to 49, and the third group consists of individuals above 50 years old. We select the six participants to ensure there are two participants in each of the three defined age groups. Their average age is 40 years (*Standard Deviation* \approx 21.26, *Median* = 30.5, *Min* = 22, *Max* = 69). In this context, 66.6% (N=4) identify as female, while 33.4% (N=2) identify as male. Furthermore, 83.33% (N=5) have completed higher education, and 16.67% (N=1) have completed high school. Additionally, 33.33% (N=2) are employed, 33.33% (N=2) are students, and 33.33% (N=2) are retired. All selected participants had prior experience with the Caixa Tem App for emergency aid usage.

Time Efficiency. Participants using the Final Design prototype require an average of 40.1 seconds for each task (*Standard Deviation* \approx 19.38 seconds, *Median* \approx 39.33 seconds, *Min* = 19 seconds, *Max* \approx 75.66 seconds). Thus, the Final Design prototype exhibits a time efficiency 10.32% higher than the Original prototype and 20.21% higher than the Redesign prototype.

The graph in Figure 13 shows the time efficiency of each prototype for each of the 10 proposed tasks. Thus, we observe that, on average the Final Design prototype requires less time compared to the other versions to complete the tasks. However, this difference varies from task to task. For example, when analyzing Task 8, which involves generating a code for cardless withdrawal, despite the Final Design prototype requiring less time than the others, the results are close, as we do not make significant changes for this functionality flow. However, when analyzing Task 5, which involves generating a QR Code to receive a Pix, we see that we could reduce the time required to complete this task compared to the other prototype versions. Coincidentally, we make the most modifications for this functionality when comparing the Final Design and Redesign prototypes.



(a) Improvement in the Balance Inquiry Function

(b) Improvement in the Pay Your Bills Function

Figure 11. Comparison of improvements made to the functionalities of balance inquiry and bill payment

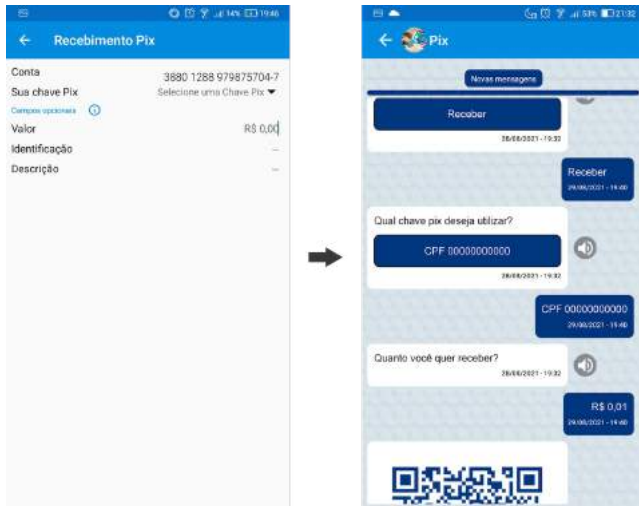


Figure 12. Presentation of before and after the enhancement in the flow of the functionality to receive a Pix via QR Code

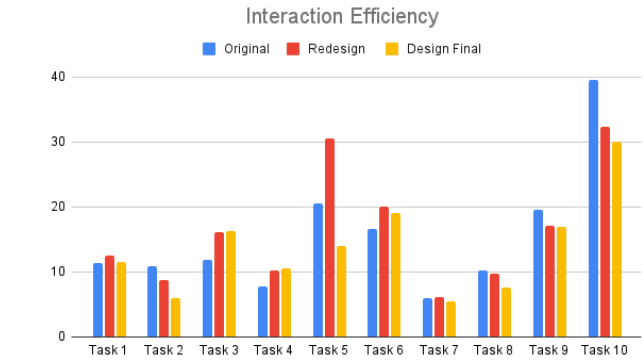


Figure 14. Interaction Efficiency Second Round Graph

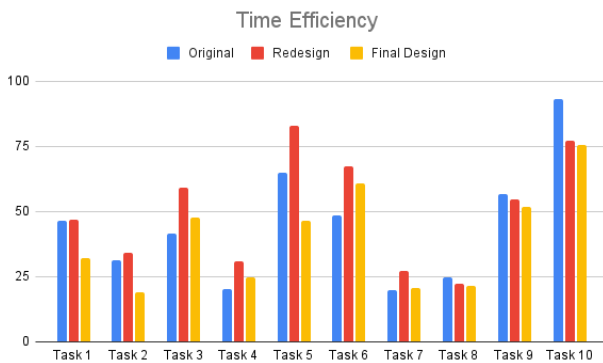


Figure 13. Time Efficiency Second Round Graph

Interaction Efficiency. Participants using the Final Design prototype require an average of 13.75 interactions for each task (*Standard Deviation* ≈ 7.37 , *Median* = 12.75, *Min* = 5.5, *Max* = 30). Thus, the Final Design prototype exhibited an interaction efficiency 10.81% higher than the Original prototype and 15.90% higher than the Redesign prototype.

The graph in Figure 14 illustrates the interaction efficiency of each prototype for each of the 10 proposed tasks in the Usability Test. Thus, we observe that, on average the Final Design prototype require a lower number of interactions with the system compared to the other versions. However, this difference varies from task to task. For example, consider Tasks 1 and 7, which involve respectively checking the account balance and consulting the payment calendar for Bolsa Família. We see that there is no significant difference in the number of interactions required because these tasks do not undergo expressive modifications that change the flow or restructure the way the functionality. On the other hand, when analyzing Task 5, which involves generating a QR Code to receive

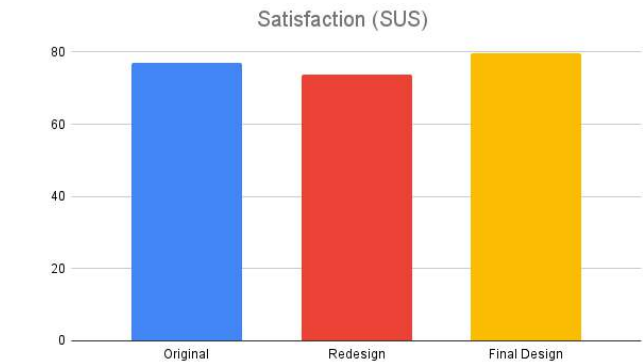


Figure 15. Second Round Satisfaction Graph

a Pix, we observe that there is a decrease in the number of interactions required considering the Final Design prototype. This happens because this task undergoes the most significant modifications in its flow, justifying the reduction in the number of interactions.

Satisfaction. Users testing the Final Design prototype rate the SUS scale with an average of 79.58 points (*Standard Deviation* ≈ 16.38 , *Median* ≈ 81.25 , *Min* = 52.5, *Max* = 100). Therefore, the Final Design prototype shows an average satisfaction 3.24% higher than the Original prototype and 7.90% higher than the Redesign prototype.

The graph in Figure 15 compares user satisfaction among prototypes. Therefore, we observe that, based on the responses obtained from the forms filled out by the participants, the Final Design prototype has a slightly higher score compared to the other prototypes.

User Experience. Figure 16 illustrates the results of the UEQ. Analyzing each scale through the t-test, we conclude that there are no significant differences in any of the six UEQ scales. Therefore, we conclude that both prototypes (Redesign and Final Design) provide an equivalent user experience according to the data collected in our tests.

Effectiveness. The Final Design prototype demonstrated

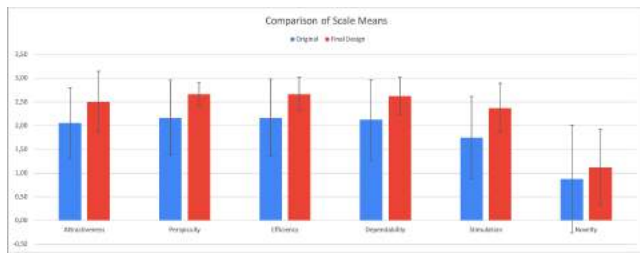


Figure 16. User Experience Graph Second Round

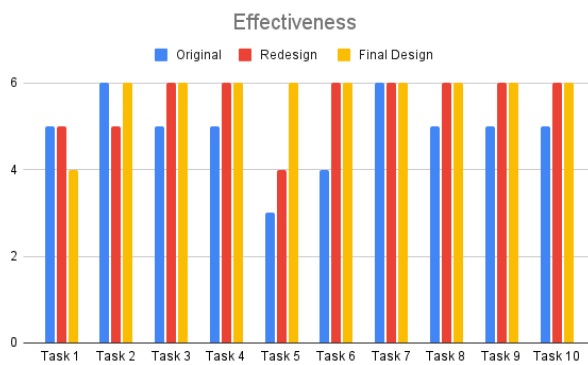


Figure 17. Comparison chart of effectiveness in each tested prototype

an effectiveness rate of 96.67%. Specifically, participants successfully completed 58 out of the 60 tasks. Thus, the Final Design prototype shows an effectiveness 18.38% higher than the Original prototype and 3.57% higher than the Redesign prototype.

The chart in Figure 17, illustrates that the Final Design prototype has better effectiveness compared to the other versions. However, there is variability in effectiveness for each task. For example, in Task 7, which involves checking the payment schedule for the Bolsa Família program, there is no variation in effectiveness across any of the prototypes. This is because we do not make any significant modifications to the flow or structure of using the functionality in any of the prototype versions. However, if we examine Task 5, which involves generating a QR Code to receive a Pix payment, we notice a significant improvement in its effectiveness in the Final Design prototype. This improvement is attributed to the fact that we make the most modifications to the flow of this task.

5.2.3 Final Questionnaire

At the end of the Usability Tests, we ask participants if they intended to use the application regularly based on their experience with the prototypes. Of the participants, 66.7% (N=12) stated affirmatively, while 33.3% (N=6) mentioned the possibility, and no participant (N=0) indicated that they would not use it. The justifications vary but could be grouped into two sets, reflecting the same percentage division in responses. Those who affirm they would use the application cited reasons such as “ease”, “convenience”, and “quick accessibility”. On the other hand, those who responded with only a maybe mentioned considerations such as “use only as needed”, “difficulty in finding some functions despite the convenience”, and “less comfortable than other banking systems due to the chat simulation”.

We also query the participants regarding their opinion on

the application’s approach of attempting to simulate a chat. The responses are unanimous: all agree that this approach indeed facilitates the use of the application. Some of them comment that it is especially helpful for users who are less familiar with banking applications, making this approach more accessible to such an audience. On the other hand, only two participants mentioned that, despite agreeing that it facilitates usage, they prefer the standard approach of other Brazilian banks like Caixa, Banco do Brasil, and NuBank because it is more straightforward. Thus, there is no need to wait for the entire automated chat flow.

5.3 Threats to Validity

In this section, we present the threats to the validity of our study. By following Wohlin et al. [Wohlin *et al.*, 2012], we organize the threats as Construct validity, Conclusion validity, External validity, and Internal validity.

Construct validity. We measure the time efficiency by means of the time in seconds. However, we do not address tiny differences of time participants can take to perform the same tasks for different prototype versions. For example, in Figure 13, participants take a few more seconds to complete Task 4 for both Redesign and Final Design than Original. In particular, the interaction flow and screen do not change between the three prototype versions for Task 4. To mitigate this threat, we conduct two rounds of Usability Tests and we also consider other quality factors such as user experience and satisfaction.

Additionally, we select 16 undergrad students to carry out the Heuristic Evaluation. Although they have had previous practice for these kind of evaluation during the Human-Computer Interaction course, we acknowledge that they are not long-term experts. This limitation might explain why we identified a few imprecise or mistaken violations. However, we analyzed all the violations and discarded those. Thus, they do not impact our results.

Conclusion validity. We consider ten Caixa Tem tasks in our Usability Tests. Although, the selected tasks are among the most used by users, it is still only a fraction of Caixa Tem’s available features. This way, it would be interesting to extend the scope of our Usability Tests to consider more tasks and consequently mitigate this issue.

External validity. In this study, we investigate usability of Caixa Tem solely. Therefore, we could not generalize our findings to other applications. Nonetheless, as we explain in Section 4.2, the Caixa Tem application is crucial for the Brazilians, especially during the COVID-19 pandemic. Thus, our results could bring better usability for this application and eventually benefit millions of users.

Internal validity. We use Figma to prototype the three Caixa Tem version: Original, Redesign, and Final Design. Although this tool is widely used for prototyping [Staiano, 2022], it has limitations. For example, we found four violations regarding it in our Heuristic Evaluation, as illustrated in Table 5. Unfortunately, we could not carry out an alternative Heuristic Evaluation with the Caixa Tem application itself to mitigate such limitations. This happens due to the sensitive data that this app handle. Therefore, people would not let we access their real Caixa Tem account.

6 Discussion

In this section, we discuss the hypotheses raised in this study in relation to the results presented in Section 5. We present our considerations by comparing the Original and Final Design prototypes. Thus, we obtain the following scenario for each of the hypotheses:

- **H1** - *The Final Design prototype provides greater user satisfaction compared to the Original prototype.* By examining the satisfaction metric in the previous section (Section 5), we can observe that indeed, the Final Design prototype achieves a System Usability Scale (SUS) score of approximately 79.58 points. This score is higher than the Original prototype, which obtained 77.03 points. Therefore, the Final Design prototype has a satisfaction level 3.24% higher than the Original prototype. Thus, we can affirm that the first hypothesis **H1** is valid in our study.
- **H2** - *The Final Design prototype requires less time, on average, for users to complete tasks compared to the Original prototype.* Upon analyzing the time efficiency metric, we observe that the Final Design prototype achieved an average time for each task of 40.1 seconds, while the Original prototype has an average time of 44.71 seconds. Consequently, there is a time efficiency improvement of 10.32% in the Final Design prototype compared to the Original prototype. Thus, we can affirm that the second hypothesis **H2** is valid in our study.
- **H3** - *The Final Design prototype, on average, requires a lower number of user interactions to complete a task compared to the Original prototype.* When analyzing the interaction efficiency metric, we observe that the Final Design prototype requires an average of 13.75 interactions to complete a task, whereas the Original prototype requires 15.41. Therefore, the Final Design prototype exhibits an interaction efficiency that is 10.81% higher than the Original prototype. Thus, we can affirm that the third hypothesis **H3** is valid in our study.
- **H4** - *The Final Design prototype provides a better user experience compared to the Original prototype.* When analyzing the user experience metric in Section 5.2.2, we notice that there is no significant difference when comparing the Original prototype with the Final Design prototype on any of the six scales. Thus, we can affirm that both experiences are equivalent based on the data collected in the study. Therefore, we can refute the fourth hypothesis **H4** in our study.
- **H5** - *The Final Design prototype ensures better effectiveness in user-performed tasks compared to the Original prototype.* When analyzing the effectiveness metric, we observe that the Final Design prototype exhibits an effectiveness of 96.67%, while the Original prototype has an effectiveness of 81.66%. Consequently, the Final Design prototype demonstrates an effectiveness that is 18.38% higher than the Original prototype. Thus, we can affirm that the fifth hypothesis **H5** is valid in our study.

Therefore, we validate four out of the five considered hy-

potheses. Hence, we only refute hypothesis **H4**, which assesses the user experience, as the data indicates that the prototypes are equivalent.

Heuristic Evaluation is essential for identifying and proposing solutions to usability issues, highlighting violations such as the lack of standardization in the iconography of the Original prototype. When analyzing the application's approach, it is possible to identify violations of the *H7 - Aesthetic and minimalist design* heuristic, leading to suggestions for creating a standardized iconography. Although effective in problem identification, Heuristic Evaluation alone does not fully address user perspectives, being limited to the evaluator's viewpoint.

In this context, it is crucial to complement Heuristic Evaluation with Usability Testing to comprehensively address usability issues. Using the example of the functionality to receive a Pix via QR Code, the violations identified and corrected in the redesign version were not sufficient, necessitating Usability Testing. These tests revealed the need for significant changes in the flow of the functionality to enhance usability in the Caixa Tem application. Therefore, the Usability Tests provided sufficient input to make improvements to the interaction in Caixa Tem. Thus, we modified the flow of this functionality, as detailed in Section 5.2.1, with the aim of improving usability at this point in the Caixa Tem application.

Users of the Final Design prototype, incorporating all modifications, take, on average, 40.1 seconds to complete each task, 4.5 seconds less than the Original prototype (44.71 seconds). Despite the 10.32% decrease in average time, in practice, this difference is not deemed significant. No user expressed complaints about the time required, considering it adequate for successful banking tasks. This metric is not critical, and in certain cases, increasing the average time may be acceptable to enhance other usability and effectiveness criteria.

Regarding the average number of interactions required to complete each task, we also observe a reduction when comparing the Original prototype to the Final Design prototype. In the Original prototype, an average of 15.41 interactions is needed for each task, whereas in the Final Design prototype, it is 13.75. Therefore, the difference is not as pronounced as with time. If we analyze the graphs of time efficiency and interaction efficiency in Section 5, we notice a certain correlation between the results for these metrics. Increasing or decreasing the time needed for a task corresponds to an increase or decrease, respectively, in the number of interactions required. Thus, we realize that the more interactions needed for a task, the longer the time required to complete that specific task.

Among the results for the selected metrics, effectiveness is the one where we achieved the best result. It is crucial for users to successfully complete the tasks they need in the application; otherwise, this audience may, for example, miss out on receiving Emergency Aid through Caixa Tem. Thus, ensuring high effectiveness is fundamental. Upon analyzing the effectiveness results obtained in the study, we observe an improvement in the Final Design prototype. In this context, it resulted in an effectiveness of 96.67%, where 58 out of the 60 tasks performed in the usability test were success-

fully completed. On the other hand, the Original prototype demonstrated an effectiveness of only 81.67%, with 49 out of the 60 tasks successfully completed.

Considering effectiveness as crucial for Caixa Tem, it is essential to address the two usability test failures that prevented 100% effectiveness in the final prototype. Both failures occurred in the same functionality and user profile: checking the balance for users above 50 years old. The difficulty arises from the distinct location of the functionality in the application header, while others are in the main menu. The attempt to correct the button text was unsuccessful, suggesting the inclusion of this functionality in the main menu as a future alternative. It is relevant to note that users below 50 years old faced fewer difficulties in this task.

It is also possible to conclude that one of the tasks that showed the most improvement in these three discussed metrics (Time Efficiency, Interaction Efficiency, and Effectiveness) was Task 5, representing the task of receiving a Pix via QR Code. In particular, we made the correction of changing the approach of the application flow, transitioning from a static screen to a chat simulation like the other functionalities of the system, as detailed in Section 5.2.1. This demonstrates that the corrections have a positive impact on these metrics, especially in terms of effectiveness, as 100% of users were able to successfully complete the task in the Final Design prototype.

When analyzing the subjective metrics in the study, we note that the Final Design prototype has a slightly higher average score on the SUS scale, registering 79.58 points, compared to the Original prototype's 77.08 points in satisfaction. Regarding the user experience, there is no significant difference in the six scales of the UEQ, suggesting that both prototypes offer equivalent experiences. We conclude that the improvements in efficiency, interaction, and effectiveness do not negatively impact user satisfaction and experience. The Original prototype, being a system in production, demonstrates good results in these metrics, indicating a likely concern with user satisfaction and experience.

7 Final Considerations

This study focuses on the usability of the Caixa Tem application, which played a vital role during the COVID-19 pandemic, being the means used by the Federal Government to provide emergency financial assistance to the population. The main goal was to identify and correct usability issues, followed by prototyping and testing the proposed improvements. Initially, we faithfully replicated the original Caixa Tem application, conducting a Heuristic Evaluation to identify and propose corrections for identified usability problems.

After analyzing the Heuristic Evaluation, we generated a new version of the prototype, implementing the suggested corrections, and conducted Usability Tests with users to assess the effectiveness of the improvements. The results indicated improvements in each version of the prototype, with slight variations in metrics such as satisfaction and user experience. It is worth noting that no modifications were made that would alter the original purpose of the application; on

the contrary, the changes aimed to enhance the Caixa Tem's base design. We particularly highlight improvements in the effectiveness metric, demonstrating that the proposed corrections contribute positively to the user experience.

Furthermore, we propose a set of specific corrections that, based on the obtained data, can be adopted by the Caixa Tem application to enhance its usability. These improvements not only address the demands of society, providing a more usable prototype but also contribute to optimizing the efficiency and effectiveness of the system in relation to the tasks performed by users.

Last but not least, as future work, we plan to conduct Heuristic Evaluations and Usability Tests for the real Caixa Tem application. This study would bring insights about real problems that may occur during usage such as network connection failures or occasional bugs. We also plan to evaluate the Caixa Tem application regarding accessibility and communicability. An assessment regarding these additional quality criteria could bring more insights for improvement that we could not detect through Heuristic Evaluation and Usability Tests. Additionally, considering more ethnographic studies for target participants could help us to better understand issues that are specific to a certain population group, such as elderly people. Finally, we intend to turn our work available so that our findings could be useful for the Caixa Tem designers.

Declarations

Availability of data and materials

The datasets generated and/or analyzed during the current study are available in <https://sites.google.com/view/jis2024>

References

- Alotaibi, M. B. (2016). Comparing the usability of m-business and m-government software in Saudi Arabia. *International Journal of Advanced Computer Science and Applications*, 7(1):117–123.
- Boren, T. and Ramey, J. (2000). Thinking aloud: Reconciling theory and practice. *IEEE transactions on professional communication*, 43(3):261–278.
- Brooke, J., Jordan, P., Thomas, B., Weerdmeester, B., and McClelland, I. (1996). *Usability evaluation in industry*. CRC Press.
- Budde, R., Kautz, K., Kuhlenkamp, K., and Züllighoven, H. (1990). What is prototyping? *Information Technology & People*, 6(2/3):89–95.
- Caixa Econômica Federal - CEF (2020). Aplicativo Caixa Tem. <https://www.caixa.gov.br/caixatem/>, Access on 16 May 2024.
- Cardoso, B. B. (2020). A implementação do auxílio emergencial como medida excepcional de proteção social. *Revista de Administração Pública*, 54:1052–1063. DOI: <https://doi.org/10.1590/0034-761220200267>.
- Chisman, J., Diller, K., and Walbridge, S. (1999). Usabil-

- ity testing: A case study. *College & research libraries*, 60(6):552–569.
- Cristóvam, J. S. d. S., Saikali, L. B., and Sousa, T. P. d. (2020). Governo digital na implementação de serviços públicos para a concretização de direitos sociais no brasil. *Sequência: estudos jurídicos e políticos*, (84):209–242. DOI: <https://doi.org/10.5007/2177-7055.2020v43n89p209>.
- Cunha, M. A. V. C. d. and Miranda, P. R. d. M. (2013). O uso de tic pelos governos: uma proposta de agenda de pesquisa a partir da produção acadêmica e da prática nacional. *Organizações & sociedade*, 20:543–566.
- Figma (2011). Figma. <https://www.figma.com/about/>, Access on 16 May 2024.
- Filho, G. K., Guerino, G. C., and Valentim, N. M. C. (2023). Usability and user experience of multi-touch systems: A systematic mapping study and benchmark. *Journal on Interactive Systems*, 14(1):292–316. DOI: 10.5753/jis.2023.3279.
- Gere, C. (2009). *Digital culture*.
- Gumussoy, C. A. (2016). Usability guideline for banking software design. *Computers in Human Behavior*, 62:277–285. DOI: <https://doi.org/10.1016/j.chb.2016.04.001>.
- Gupta, M., Mehta, D., Punj, A., and Thies, I. M. (2022). Sophistication with limitation: Understanding smartphone usage by emergent users in india. In *Proceedings of the 5th ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies*, page 386–400. DOI: <https://doi.org/10.1145/3530190.3534824>.
- Hasan, L. (2013). Heuristic evaluation of three jordanian university websites. *Informatics in Education*, 12(2):231–21.
- IBGE (2018). Pesquisa nacional por amostra de domicílios contínua. <https://www.ibge.gov.br/estatisticas/sociais/trabalho/9171-pesquisa-nacional-por-amostra-de-domicilios-continua-mensal.html?#t=destaques>, Access on 16 May 2024.
- IBGE (2019). PNS - pesquisa nacional de saúde. <https://www.ibge.gov.br/estatisticas/sociais/saude/9160-pesquisa-nacional-de-saude.html?#t=resultados>, Access on 16 May 2024.
- Jardim, J. M. (2000). Capacidade governativa, informação e governo eletrônico. *DataGramaZero – Revista de Ciência da Informação*, 1(5).
- Kosakowski, J. (1998). *The benefits of information technology*. ERIC Digest.
- Laugwitz, B., Held, T., and Schrepp, M. (2008). Construction and evaluation of a user experience questionnaire. In *Symposium of the Austrian HCI and usability engineering group*, pages 63–76.
- Lewis, J. R. (2006). *Usability testing*. John Wiley Sons.
- Lichter, H., Schneider-Hufschmidt, M., and Zullighoven, H. (1994). Prototyping in industrial software projects-bridging the gap between theory and practice. *IEEE transactions on software engineering*, 20(11):825–832.
- Lisbôa, D., Rocha, T. d., Machado, L., Caldeira, C., and de Souza, C. (2021). Working in the covid-19 pandemic: an observational study. *Journal on Interactive Systems*, 12(1):283–293. DOI: 10.5753/jis.2021.2178.
- Marchionini, G., Samet, H., and Brandt, L. (2003). Digital government. *Communications of the ACM*, 46(1):25–27.
- Marky, K., Zimmermann, V., Funk, M., Daubert, J., Bleck, K., and Mühlhäuser, M. (2020). Improving the usability and ux of the swiss internet voting interface. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13. DOI: <https://doi.org/10.1145/3313831.3376769>.
- Monteiro, I. T., Lourenço, E. L. d. Q., Brilhante, M. Q. d. L., de Freitas, A. T., Oliveira, F. C. d. M. B., de Castro, F. E. B., and de Oliveira, A. C. B. (2022). Design and evaluation of a prototype of a children’s educational application during and for the covid-19 pandemic and beyond. *Journal on Interactive Systems*, 13(1):54–76. DOI: 10.5753/jis.2022.2004.
- Monteiro, M. d. S., Batista, G. O. d. S., and Salgado, L. C. d. C. (2023). Investigating usability pitfalls in brazilian and foreign governmental chatbots. *Journal on Interactive Systems*, 14(1):331–340. DOI: 10.5753/jis.2023.3104.
- Motta, P. R. F. (2003). *Agências reguladoras*. Editora Manole Ltda.
- Nielsen, J. (1994). *Usability engineering*. Morgan Kaufmann.
- Nielsen, J. (1995). How to conduct a heuristic evaluation. *Nielsen Norman Group*, 1(1):8.
- Nielsen, J. (2005). Ten usability heuristics. <https://www.nngroup.com/articles/ten-usability-heuristics/>, Access on 16 May 2024.
- O’neill, J., Dhareshwar, A., and Muralidhar, S. (2017). Working digital money into a cash economy: The collaborative work of loan payment. *Computer Supported Cooperative Work: CSCW: An International Journal*, 26:1–36. DOI: <https://doi.org/10.1007/s10606-017-9289-6>.
- Riihioho, S. (2018). Usability testing. *The Wiley Handbook of Human Computer Interaction*, 1:255–275.
- Rudd, J., Stern, K., and Isensee, S. (1996). Low vs. high-fidelity prototyping debate. *interactions*, 3(1):76–85.
- Sahasrabudhe, S. and Lockley, M. (2014). Understanding blind user’s accessibility and usability problems in the context of myitlab simulated environment. In *Proceedings of the 20th Americas Conference on Information Systems*, pages 1–14.
- Staiano, F. (2022). *Designing and Prototyping Interfaces with Figma: Learn essential UX/UI design principles by creating interactive prototypes for mobile, tablet, and desktop*. Packt Publishing Ltd.
- Viana, A. C. A. (2020). Aplicativo utilizado para cadastro do auxílio emergencial pode ser excludente.
- Vieira and Andrade (2024). Online appendix. <https://sites.google.com/view/jis2024>.
- Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., and Wesslén, A. (2012). *Experimentation in software engineering*. Springer Science Business Media.