

**Analysis of Non-Functional Requirements Based on Eye Tracking**

**By Li Rongchen**

**Supervisor: Li Tong**

**Submitted to the Fan Gongxiu Honors College, BJUT**

**in Partial Fulfillment of Requirements for**

**the Degree of Bachelor of Engineering in Computer Science**

**Beijing University of Technology**

**Beijing, China**

**June, 2021**

# Abstract

Quality requirements, also known as non-functional requirements, are generally harder to express measurably, making them more challenging to analyze. Particularly, non-functionals tend to be properties of a system as a whole and cannot be verified for individual components. It is even more difficult because it is often not explicitly mentioned in the official requirements specifications and miss at the beginning of system development. As a result, detection and modification of these functions in the subsequent maintenance of the system cause high costs. Constantly evolving systems rely heavily on user feedback to assess how non-functional requirements are met. It analyzes explicit user feedback, such as App Store ratings and reviews. However, explicitly user feedback has challenging to collect and is therefore considered time-consuming. Many users are reluctant to provide such explicit feedback. Additionally, analysis of the explicit user feedback also has flaws, making it challenging to accurately define their requirements, resulting in lower efficiency and Accuracy in assessing the degree to which the needs are being met. A potential solution is to parse the users' implicit comments from the user's usage process, which does not require an additional load on the user. Moreover, users' implicit reactions come directly from users' objective reflections rather than their subjective expressions, which is more reliable.

This thesis investigated a new method of non-functional requirement analysis, which investigates eye movement patterns to collect users' implicit feedback automatically. Specifically, we establish semantic connection between several eye-movement metrics and user perceptions based on cognitive and empirical evidence from existing studies, summarize their correspondences with six widely-used non-functional requirements, and attempt to formalize several eye-movement patterns. Finally, the correctness of the correspondences is verified by designing and practicing experiments, and six quantitative evaluation criteria for these six non-functional requirements are proposed. The method approach can thus systematically assess the satisfaction of six non-functional requirements, quietly unveiling users' needs for the software they use during normal user usage or user testing of the software they use. Further, the proposed provides the details of the precise location implementation that do not satisfy the non-functional requirements and reduces the cost of fixes and enhancements to the system. Thanks to our modular design and implementation of the “Eye-tracking Based Requirements Assessment Framework”, the tool is now quite useful. Existing applications can quickly start using our approach to assess the level of non-functional requirement fulfillment of their applications by plugging in inspection code.

**Keywords: non-functional requirements ; eye movement pattern; empirical evaluation ; eye-tracking**

**Contents**

[Abstract III](#_Toc105867767)

[**Chapter 1 : Introduction** 1](#_Toc105867768)

[1.1 Background 1](#_Toc105867769)

[1.2 Significant of the study 2](#_Toc105867770)

[1.3 Objectives 3](#_Toc105867771)

[1.4 Thesis outline 4](#_Toc105867772)

[**Chapter 2 : Related Work** 5](#_Toc105867773)

[2.1 Analysis based on explicit user feedback 5](#_Toc105867774)

[2.2 Analysis based on implicit user feedback 5](#_Toc105867775)

[2.3 Eye tracking in software engineering 6](#_Toc105867776)

[2.4 Definitions of Non-functional requirements 7](#_Toc105867777)

[**Chapter 3 : Eye-tracking Based Requirements Assessment Framework** 8](#_Toc105867778)

[3.1 Obtaining Eye-tracking Metrics 9](#_Toc105867779)

[3.1.1 First-level metrics 9](#_Toc105867780)

[3.1.2 Second-level metrics 9](#_Toc105867781)

[3.1.3 Third-level metrics 9](#_Toc105867782)

[3.1.4 Fourth-level metrics 10](#_Toc105867783)

[3.2 Analysis based on Metrics 11](#_Toc105867784)

[3.2.1 Accuracy 12](#_Toc105867785)

[3.2.2 Robustness 13](#_Toc105867786)

[3.2.3 Consistency 15](#_Toc105867787)

[3.2.4 Ease of use 16](#_Toc105867788)

[3.2.5 Device efficiency 17](#_Toc105867789)

[3.2.6 Accessibility 19](#_Toc105867790)

[3.3 Analysis based on Eye Movement Patterns 20](#_Toc105867791)

[**Chapter 4 : Evaluation** 22](#_Toc105867792)

[4.1 Selection of Participants 22](#_Toc105867793)

[4.2 Experimental Materials and Tasks 22](#_Toc105867794)

[4.3 Experimental Design 32](#_Toc105867795)

[4.4 Results 32](#_Toc105867796)

[4.4.1 Accuracy 32](#_Toc105867797)

[4.4.2 Robustness 35](#_Toc105867798)

[4.4.3 Consistency 38](#_Toc105867799)

[4.4.4 Ease of use 41](#_Toc105867800)

[4.4.5 Device efficiency 44](#_Toc105867801)

[4.4.6 Accessibility 47](#_Toc105867802)

[4.5 Real-World Test 50](#_Toc105867803)

[**Chapter 5 : Conclusions and Future Work** 53](#_Toc105867804)

[**Bibliography** 55](#_Toc105867805)

[**Acknowledgements** 59](#_Toc105867806)

**Chapter 1 : Introduction**

## 1.1 Background

Efficiently and precisely collecting user feedback is essential for understanding user requirements, based on which software systems evolve continuously. User feedback contains both functional requirements (FR) and non-functional requirements (NFR), with NFRs being more valuable. Roughly speaking, FRs describe what the system should do, while NFRs limit how these FRs can be implemented [1]. NFRs are widely recognized as playing a pivotal role in software development. Unlike FRs, which often have structured methods to capture them, NFRs are challenging to summarize and extract. As a result, NFRs are often not well understood and fully considered in the software development process [2].

For a long time, much of the attention in software engineering has been focused on the functional notations and techniques used to define and deliver the software systems that must perform. The reality is that the problems are more non-functional than functionally oriented, e.g., low productivity, slow processing, prohibitive costs, low quality, and unsatisfied customers [3]. In software engineering delivery, customers need usable applications that they can operate immediately with minimal training and appropriate user interface design to minimize errors. It is generally accepted that omissions or commissioning errors in the development and proper consideration of these requirements are the most expensive and difficult to correct once the information system is complete [4].

In recent decades, treating quality features as a whole, rather than just as functions, has been the focus of work in the field of goal-oriented requirements engineering. In particular treats non-functionality at a high level of abstraction for both the problem and the solution has been widely defined by the NFR Framework [3]. The focus of recent research has shifted from defining and abstracting NFR systems to finding ways to assess the extent to which the NFR is being met from actual production use cases.

Currently, the mainstream approach to assessing requirement satisfaction from 'users' feedback relies mainly on explicit user feedback, such as app store scores and reviews. Researchers sort and filter this predictive information to derive useful information from it [5–7]. However, these methods rely on active user feedback, thus missing the needs of users who are reluctant or not good at giving feedback.

In addition, studies in cognitive science showed that the 'participants' perception of their behavior does not always agree with their underlying processes and intentions [8] especially when they have to detach from actual usage scenarios to make feedback. Some research has also begun to focus on methods for implicitly capturing user needs [9,10]. These studies use conditional random field-based desire models to induce user NFRs. They require careful granularity adjustment by relevant domain experts before use and the results are reasoned by domain experts based on subjective judgment. These studies use conditional random field-based desire models to induce user NFRs. They require careful granularity adjustment by relevant domain experts before use and the results are reasoned by domain experts based on subjective judgment. Their analysis is complex and tedious, requiring a high threshold for the average engineer to obtain useful information from them. Moreover, these methods are difficult to migrate between systems and especially difficult to deploy in practical applications. In addition, subsequent research to complement and extend them in their theoretical systems faces a greater challenge.

## 1.2 Significant of the study

To address these challenges, we propose an eye-tracking-based requirements assessment framework.

The thesis contributes to the state of the arts in the following three aspects.

* Firstly, instead of the missing and inaccuracy in the explicit user feedback analyses approach, the thesis proposes quantitatively assessing the satisfaction of NFRs in a non-interruptive manner, collect user feedback implicitly using eye-tracking techniques.
* Secondly, the theoretical model is divided into two steps: “Mapping NFRs satisfaction level to user behavior” and “Mapping user behavior to eye movement index”. Each step in this section has a clear reasoning process that is highly interpretable and easily understood by others.
* Thirdly, the model is easy to apply and has very clear application scenarios. With simple hardware requirements (Infrared camera shown in Figure 1) and simple software usage (modular plug-in framework), the model can be applied to the user testing phase of software development so that developers can quickly complete usability testing and reduce the time and labor costs of the process. The model analyzes where and why NFRs are not being satisfied at a very granular level of granularity, which makes it easy for developers of existing applications to maintain and modify.

图形用户界面

描述已自动生成

Figure 1 Hardware required for the project and practical usage scenarios

## 1.3 Objectives

This thesis will accomplish the following three objectives

* Firstly, we need to summarize the mapping relationships from the existing studies of users' eye-tracking behaviors to the corresponding eye-movement metrics. We investigate and study the existing metrics of eye-movement data, based on which we establish a set of metrics that are closely related to the assessment of NFRs.
* Secondly, we need to identify the mapping relationship from non-functional requirements to users' eye-tracking behaviors, which is an unstudied area, and for this purpose we empirically as well as investigate and establish the semantic links between six typical NFRs and the above metrics.
* Finally, we propose an evaluation plan. The validity of the theoretical model is verified by practical experiments.

## 1.4 Thesis outline

In Chapter 1, we present the background of this study when it was proposed and highlight the significance of this study and the clear objectives that need to be accomplished. In Chapter 2, we will present in detail the research methodology of similar studies in the field of requirements analysis and give the definitions of the six NFRs used in this study from past studies. In Chapter 3, we will formally present the derivation of the theoretical model of the Eye-tracking Based Requirements Assessment Framework, following the actual implementation process of the framework. The association between user behavior and eye-tracking metrics is given first, and then we explain how we build the mapping from NFR to user behavior and then to eye-tracking metrics. In Chapter 4, we present a detailed experimental plan to verify the correctness of the theoretical model, then pend parameters of the model determined by the experimental results. The validity of the model is verified by simulating a test scenario with real applications. Finally, in Chapter 5, we conclude and discuss this study and provide an outlook on how to proceed with the subsequent work.

**Chapter 2 : Related Work**

Obtaining users' feedback can be classified into two categories: one is to get explicit feedback through their reviews, and the other is to get implicit feedback through their behavior. The following sub-headings review the most important related works based on this classification as follows:

## 2.1 Analysis based on explicit user feedback

Reviews are the easiest way to obtain explicit user feedback about the application, which contains much requirement-related information, such as bug reports, feature requests, user experiences, ratings, etc. [5]. User reviews are characterized by their huge number. A recent study found that mobile apps receive about 23 user reviews per day, while popular apps, such as Facebook, receive an average of 4275 user reviews per day [11]. Several kinds of research are currently focused on identifying valuable information from unstructured and informal user comments.

Various classification techniques have been developed to distinguish between types of reviews [5–7], most of which focus on functional requirements. However, user reviews include both functional and non-functional requirements, and some studies focus on finding NFRs associated with quality characteristics from user reviews [12]. Analyzing user requirements for a system from explicit user feedback faces an unavoidable problem. That is, users must actively express their requirements outside of product use. This explicit feedback usually requires additional effort or time, leaving developers at risk of missing users' perceptions who are reluctant or inept at providing feedback. In addition, user feedback may be inauthentic or distorted because it is detached from the use of the product in the field.

## 2.2 Analysis based on implicit user feedback

The relative benefit of evaluating user needs implicitly is that it does not require the user to do anything, and everything captured is relatively objective and valid. Some studies have attempted implicit NFR evaluation to explore system-user interactions quantitatively through conditional random fields to uncover potential user needs or requirements [10]; or infer human needs by monitoring environmental context and human behavioral context [9]. However, all these efforts are still exploratory and face problems such as not being easy to deploy and difficult to scale. Accordingly, a simple, straightforward, and systematic way to evaluate user requirements implicitly is much needed.

## 2.3 Eye tracking in software engineering

Eye tracking involves collecting a participant’s overt visual attention by recording eye gaze data [13,14]. Visual attention triggers the cognitive processes required for comprehension and problem solving, while cognitive processes guide visual attention to specific locations. Therefore, eye tracking is useful to study the participant’s cognitive processes and effort while performing software engineering tasks [14].

Eye trackers provide an objective, real-time, quantitative measure of eye gaze, without conscious filtering. They help researchers to study processes and intentions that participants cannot articulate [15]. An eye tracker provides additional insights into what participants were doing and why based on where they focused their attention during a task.

Eye trackers help researchers determine (1) why participants have problems finishing a task, (2) where participants expect to find certain elements, (3) whether elements are distracting, (4) how efficiently a design, layout or artifact guides participants through a task,(5) whether there are differences in the participants’ efficiency, based on their demographics or expertise, and (6) whether participants focus on details or briefly scanned the stimuli [16,17]. All of this is done objectively while the task is being executed.

However, using eye trackers correctly is vital. Conducting an eye tracking study requires dedication to fine details to make sure the data is collected correctly and accurately. The collected data must be analyzed carefully to relate participants’ fixations with their cognitive processes and intentions [18,19]. In particular, currently, there is no absolute way of knowing whether participants indeed understood parts of the stimuli on which they fixated [20].

## 2.4 Definitions of Non-functional requirements

It is well known that non-functional requirements encompass all aspects of software design and are more difficult to define, measure, test, and track than traditional functional requirements [21]. Considering the need to extract non-functional requirements related to user experience during user testing, we did not select non-functional requirements that were only relevant to software maintenance and subsequent development in the initial phase. After sifting and sorting through the non-functional requirements, we finally selected several non-functional requirements that are widely used, well-defined, closely related to user experience, and have clear characteristics [22]. The requirements for reservations are listed below, and the descriptions are taken from publicly available definitions [23].

**Accuracy**: Degree of precision to which service provides the right outcomes or effects.

**Robustness**: Degree represents the ability of the service to act properly, even if some input parameters are missing or incorrect.

**Ease of use**: Degree to which a service is effortless for users to operate and control.

**Consistency**: Degree to which representative services offer the same design pattern.

**Device efficiency**: Degree of promptness of the service in using a certain number of resources.

**Accessibility**: Degree to which people with distinctive features and capabilities can use a service to achieve a specified goal in a context of use.

**Chapter 3 : Eye-tracking Based Requirements Assessment Framework**

The structure of the entire framework is shown in Figure 2. The inputs to our framework are subtasks of the function that the developer has labeled with the key “area of interest” (AOI). In this section, we will first describe how we obtain eye-tracking metrics. Then, it is explained how we can use eye-tracking metrics to evaluate the level of satisfaction of non-functional requirements for subtasks. Finally, a method is discussed to extract several patterns from the metrics and use the patterns to more finely and comprehensively detect the user's invisible needs and evaluate the degree of satisfaction of the non-functional needs of the software.

图示

中度可信度描述已自动生成

Figure 2 The structure of the Eye-tracking based requirements assessment framework.

## 3.1 Obtaining Eye-tracking Metrics

By referring to the research methods of other eye-tracking projects [24,25], we can use several eye-tracking metrics to portray different aspects of user perception.

High-level metrics are based on low-level metrics, which require more complex processing, including but not limited to computing, de-noising, etc. The higher the level, the better the representation of narrow semantics and clear characteristics. the higher weights are given to higher-level metrics in our studies.

### 3.1.1 First-level metrics

The raw data that most affordable eye-tracking devices can collect is the Eye gaze point position on screen (X, Y). Generally, the collection of this data relies on computer vision technology to track the angle of rotation of the user's head and eyes.

### 3.1.2 Second-level metrics

The eye tracker can distinguish between Fixations and Saccades by using an event detection algorithm to locate the eye's indicated landing point. Fixations tend to mean that the user is creating awareness of a specific object, while saccades mean that the user's eyes are simply moving quickly along the position. We chose the discrete threshold-based method with excellent accuracy and Robustness [26].

### 3.1.3 Third-level metrics

Area of interest (AOI) is introduced at this level as a tool to mark stimulus areas and extract indicators specifically for these areas. We use the components in the software as the AOI area in the study.

Fixation duration (FD): Time spent gazing at the key AOI. Previous studies have used this metric to characterize the participant's effort [20]. The metric's value is related to the amount of effort the user puts into the AOI.

Fixation count (FC): The number of fixations in key AOI. It can also describe the amount of energy or effort that participant spends on the AOI. However, prolonged gaze may produce only a single or a small number of fixation counts, unlike fixation duration.

Fixation rate (FR): Ratio of the total number of fixations on key AOI to others. A lower fixation rate indicates a lower efficiency in search tasks: participants spend more effort to find relevant areas [27]. Higher rates may indicate that more effort is required to complete tasks [20].

First fixation time (FFT): Time to the first fixation in an AOI can indicate the key AOI's attention-grabbing level.

### 3.1.4 Fourth-level metrics

The concept of “scan path”, is introduced in the four-level metrics, a sequence of gaze points or AOIs that describes the length and duration of fixation. Performing comparisons between scan paths allows identification and analysis of ' 'participant's viewing strategies for solving tasks [28].

Scan path accuracy (SPA): The ratio of the number of key AOIs to the number of other AOIs among the AOIs that have been gazed at. A higher value indicates a higher level of understanding of the task by the participant [29].

Edit distance (ED): The editing cost of converting a scan path to the shortest path determined by the developer is calculated by basic operations such as insert, replace and delete. The smaller the value, the more the participant's behavior matches the developer's expectation.

Number of attention switches (AS): The total number of switches between a list of AOIs per unit time. A higher number of switches means that participants are more uncertain about the task being performed.

A convex-hull area (CA): The smallest convex set of fixations that contains all of the participant's fixation. Smaller indicates that the closer the distance between the gaze points, the less effort participant put into finding the relevant region [30].

## 3.2 Analysis based on Metrics

It is well known that non-functional requirements encompass all aspects of software design and are more difficult to define, measure, test and track than traditional functional requirements. After sifting and sorting through the non-functional requirements, we finally selected several non-functional requirements that are widely used [23], closely related to user experience and have clear characteristics. We found that once an NFR was not satisfied, some characteristic change in the user's metrics occurred. We corresponded the eye-tracking metrics to the NFRs based on the behavioral characteristics of the individual metrics that are widely accepted to respond to users as mentioned in paragraph A above, combined with the characteristics of users that appear in human-computer interaction when the individual NFRs are not satisfied. Among them, some metrics respond substantially when an NFR is not satisfied, while others do not differ significantly when this NFR is satisfied or not. The correspondence between the NFRs and the changed characteristics we have identified so far is shown in Figure 3.

图示

描述已自动生成

Figure 3 Association between NFRs and eye movement indicators

### 3.2.1 Accuracy

The accuracy is not met indicates that the service provides inaccurate results and therefore does not easily allow the user to obtain the correct meaning of the returned results. That is, the user needs to compare elements on the interface to gather more information to help them understand. Users generate search behavior to gather information about the elements on the entire interface. The search behavior causes fixation to be scattered over other AOIs, which reduces the fixation rate and the accuracy of the scan path. Therefore, the unpredictability of the user's search behavior also leads to a significant increase in edit distance, and the frequent switching of AOIs increases the number of attention switches. So, in a nutshell, when compared to the standard behavior, the user's behavior exhibits the following characteristics when inconsistent.

* Low fixation rate
* Low scan path accuracy
* Long edit distance
* High number of attention switches

However, because the user can find the relevant AOI effortlessly, the first fixation time does not change. In addition, the convex hull area is small in the search because the user is confused by the confusing returned results and is clear about the general direction. So the user will search around the key AOI, and the range will not be vast. In addition, once the key AOI has been found, the user faces no more, so his fixation count and fixation duration are as usual.

* First fixation time on relevant AOI
* Fixation count
* Fixation duration
* Convex-hull area

In summary, only fixation rate, scan path accuracy, edit distance, and the number of attention switches changed. For edit distance and number of attention switches, higher values indicate higher dissatisfaction with this NFR; for the two parameters fixation rate and scan path accuracy, lower values indicate higher dissatisfaction with this NFR. We used the following formula to assess the level of dissatisfaction with accuracy. The coefficient assigned to the fixation rate is relatively small because it is a lower-level indicator. In contrast, the coefficients assigned to scan path accuracy, edit distance, and the number of attention switches are higher because they are higher-level metrics.

### 3.2.2 Robustness

When the accuracy is not met, it indicates that the system does not work properly when the parameters entered into the system are missing or incorrect, i.e., the user's actions lead to a result that he did not expect. However, the user often easily deduces the cause of the system anomaly and will improve it in the next attempt. The retry behavior causes the user to operate the relevant AOI again, which results in double or more fixation count and fixation duration compared to the standard behavior, and the fixation rate is increased because the user does not put as much effort into the non-relevant AOI. So, when compared to the standard behavior, the user's behavior will exhibit the following characteristics when not conforming to consistency.

* High fixation count
* High fixation duration

However, the focus on and use of the relevant AOI is essentially correct until the wrong parameters are entered. It is not difficult for the user to find out what is the relevant region for completing the task, which would result in the first fixation time on relevant AOI being exactly the same as the standard behavior. The retry behavior occurs after the first correct attempt, which means that there is no significant increase in edit distance. The retry operation is concentrated in the relevant AOI and does not switch between non-relevant and relevant AOIs, which will result in the fixation rate, number of attention transitions and convex hull area not being significantly different from the standard behavior. Therefore, the following indicators will be exactly in line with the standard behavior and will not change.

* Fixation rate
* First fixation time on relevant AOI
* Edit distance
* Number of attention switches
* Convex-hull area

In summary, only fixation count and fixation duration changed. For the two parameters, fixation count and fixation duration, lower values indicate higher dissatisfaction with this NFR. We used the following equation to evaluate the level of dissatisfaction with robustness. Because fixation count and fixation duration belong to the same level of eye-tracking metrics, there is no need to add weights used to classify different levels of eye-tracking metrics.

### 3.2.3 Consistency

When Consistency is not met, it indicates that the system does not provide the same design pattern, meaning that elements with the same ideograms behave differently or that elements with different ideograms have the same appearance or behavior patterns. This can be confusing for users who have difficulty spotting the element they need in the first place. The most distinctive feature of elements that cannot be found in the first place is that the first fixation on the relevant element is late, as the user is first attracted to other elements. In addition, the user will exhibit search characteristics when forced to find the correct element of interest after a failed attempt. The act of searching causes fixation to be scattered across other AOIs, resulting in a lower fixation rate and scan path accuracy. The unpredictability of user search behavior also leads to a significant increase in edit distance; frequent switching in AOIs increases the number of attention switching. In general, when not conforming to the standard behavior, the user's behavior will exhibit the following characteristics compared to the standard behavior.

* Low fixation rate
* Late first fixation time on relevant AOI
* Low scan path accuracy
* Long edit distance
* High number of attention switches

However, in the search, the convex hull area is small because the user is just confused by the confusing design patterns and is clear about the general direction. Therefore, the user will search around the key AOI, and the range will not be very large. In addition, once the key AOI has been found, the user faces no more difficulties with the operation. So his fixation count and fixation duration are as usual. Thus, there will be no change with the following indicators that will align with the standard behavior.

* Fixation count
* Fixation duration
* Convex-hull area

To summarize, fixation rate, first fixation time, scan path accuracy, edit distance, and the number of attention switches vary depending on how well the consistency is met. The worse the level of satisfaction, the lower the fixation rate, first fixation time, and scan path accuracy followed, and the higher the editing distance and the number of attention switches followed. Therefore, we use the following formula to evaluate the satisfaction of this NFR. Note that we have set different weights for metrics at different levels, with the higher the level, the higher the weighting.

### 3.2.4 Ease of use

If Ease of use is not satisfactory, the service is hard for users to operate and control. Too much difficulty in operation leads to more effort and makes users skeptical and make multiple attempts. Too much effort and repeated attempts by the user will result in an unusually high amount of time and attention spent on key AOIs, so the fixation duration is long. Then, the task is so difficult to perform that the user may wonder whether the right action is the right action, thus triggering suspicion and a search for other elements around, which brings low scan path accuracy and a high number of attention switches. The fixation count increases because the search behavior is interspersed with the operation. So, in general, when not conforming to the standard behavior, the user's behavior will exhibit the following characteristics when compared to the standard behavior.

* High fixation count
* High fixation duration
* Low scan path accuracy
* High number of attention switches

However, because the number of fixations inside and outside the key AOIs is both large, the fixation rate is not significantly higher, and the search area is around key AOIs. Hence, the convex-hull area is small. In addition, all search behavior comes after the user has found the relevant elements which difficult to manipulate. The user is correctly directed to the relevant element to start the action, so the first fixation is early as normal, and the editing distance is short.

* Fixation rate
* First fixation time on relevant AOI
* Edit distance
* Convex-hull area

Only the fixation count, fixation duration, scan path accuracy and the number of attention switches have changed. For the fixation duration and number of attention switch parameters, lower values indicate better satisfaction of this NFR; for the scan path accuracy, higher values indicate better satisfaction of the NFR. We use the following formula to evaluate the degree of satisfaction with ease of use. The coefficient assigned to fixation duration is relatively small because it is a lower-level metric. In contrast, the coefficients assigned to scan path accuracy and attention switch count are higher because they are higher-level metrics.

### 3.2.5 Device efficiency

When Device efficiency is not met, it indicates that the service is slow or delayed in responding when a certain amount of resources are used, meaning that the user has to wait for a response to some content. The user generates a waiting behavior. The waiting is generally divided into two steps. In the first step, the user first tries to look directly at the content in the relevant AOI in order to capture the change in its content in the first place, which will lead to a very high fixation count and a very long fixation duration in the AOI. once the waiting time exceeds a certain limit, the user will carry out the second waiting behavior because of boredom, which is “start to find things for themselves”, i.e., search aimlessly in the interface. This is reflected in the eye-tracking data metrics as a significant decrease in scan path accuracy and a significant increase in the number of attention switches. Further, because the search is aimless, the convex hull area increases. So, in general, compared to the standard behavior, when inconsistent, the user's behavior exhibits the following characteristics.

* High fixation count
* High fixation duration
* Low scan path accuracy
* High number of attention switches
* Large convex hull area

However, because the first and second wait operations increase the fixation count on relevant AOI and irrelevant AOI, respectively, the fixation rate tends to be flattened, i.e., it is not too strongly differentiated from the standard behavior. In addition, the response delay due to the inefficiency of the device does not affect the user's perception of the page elements, so the first fixation time on relevant AOI is also not significantly different from the standard behavior. Ultimately, the waiting behaviors all occur after the user action is completed, so the editing distance does not change much either. The following metrics will match the standard behavior exactly and will not change.

* Fixation rate
* First fixation time on relevant AOI
* Edit distance

In summary, only fixation count, fixation duration, scan path accuracy, attention switches, and convex hull area changed. For fixation count, fixation duration, attention switches and convex hull area, higher values indicate higher dissatisfaction with this NFR; for scan path accuracy, lower values indicate higher dissatisfaction with this NFR. We use the following equation to evaluate the degree of dissatisfaction with Device efficiency. The coefficient assigned to fixation count and fixation duration is relatively small because it is a lower-level metric. In contrast, the coefficients assigned to scan path accuracy, attention switches, and convex hull area are higher because they are higher-level metrics.

### 3.2.6 Accessibility

When Accessibility is not satisfied, it indicates that the user does not have the relevant skills to access the content, meaning that the user needs a lot of searching, understanding, and experimentation to find a solution to the task gradually. However, since there are often more non-relevant AOIs than relevant AOIs on the page and the user does not have a clear understanding of the destination, the final fixation rate is still very low. The first fixation time on relevant AOIs is often late. In addition, extensive and random attempts and searches also lead to very low scan path accuracy, very long edit distances, a particularly high number of attention switches, and a convex-hull area. So, in general, compared to standard behavior, user behavior exhibits the following characteristics when inconsistency is not met.

High fixation count

* Long fixation duration
* Low fixation rate
* Late first fixation time on relevant AOI
* Low scan path accuracy
* Long edit distance
* High number of attention switches
* Large convex-hull area

In summary, fixation count, fixation duration, fixation rate, first fixation time on relevant AOI, scan path accuracy, edit distance, number of attention switches, and convex-hull area have been changed. For fixation count, fixation duration, first fixation time on relevant AOI, edit distance, number of attention switches and convex-hull area, higher values indicate higher dissatisfaction with the NFR. For the parameters fixation rate and scan path accuracy, lower values indicate higher dissatisfaction with the NFR. We use the following formula to evaluate the dissatisfaction level of Accessibility. The coefficients assigned to the four parameters fixation count, fixation duration, fixation rate, and first fixation time on relevant AOI are relatively small because they are lower-level indicators. In contrast, the coefficients assigned to the four parameters edit distance, number of attention switches, and convex-hull area are higher because they are higher-level metrics.

## 3.3 Analysis based on Eye Movement Patterns

The analysis above demonstrates that many fixed patterns recur across multiple NFRs. We carry out further work to abstract these patterns from the eye-movement metrics to help us better evaluate non-functional requirements. We use the idea of a variable-length sliding window to analyze the multiple metrics mentioned above at a finer granularity on the timeline of the user's task completion.

From the point of view of user perception, the user's behavior is always made up of multiple patterns when each NFR is not satisfied [31]. We take two patterns as an example, one for “searching” and one for “performing task”, which we will describe in more detail below. These two models have set the stage for more models to be proposed in the same way to provide a more granular analysis of the NFR assessment.

**Searching:**

Through the analysis experience, we used a low fixation rate, low scan path accuracy, and a high number of attention switches as a pattern. When they appear in combination in a window of time, they represent a search behavior used by the user, as the fixation landing point is bound to wander between multiple AOIs when searching.

There are multiple NFRs that users exhibit search behavior during task execution when they are not satisfied. For example, when the ease of use is not satisfied, the user intersperses the search behavior in the middle of the task execution to exclude the possibility of the task being difficult due to his actions. When consistency is not satisfied, users exhibit search behavior at the beginning to find the key AOI; when device efficiency is not satisfied, the user starts searching for changes on the whole page due to boredom of waiting, etc.

The searching pattern appears in different positions in the different cases in that each NFR is not satisfied, which means different for each NFR. For example, in the case of consistency, the search pattern is the core behavioral feature. Using the captured data, we can further analyze which elements conflict with the search object's key AOI design pattern, which confuses the user. However, in the case of device efficiency, the search is just a by-product when the NFR is not satisfied, so we do not need to care about the specific object.

图表

描述已自动生成

Figure 4 The raw eye-tracking data of several patterns, the yellow fluorescent area represents the critical AOI area, and the dot represents fixation. The longer the fixation duration, the larger the dot.

**Performing:**

When the fixation rate is extremely high, the fixation duration is extremely long, the scan path accuracy is extremely high, and the convex-hull area is extremely small in a period of time window, we can conclude that the user is completing the sub task's goal. Regardless of the existence of NFR unsatisfaction, the execution task pattern will always occur because, for each subtask, completion is a mandatory path to the next subtask.

When either consistency or ease of use is not satisfied, the user can only have two behavioral patterns, searching and performing. However, the difference is in the order of the patterns: the user behavior when consistency is not satisfied is “Searching” – “Performing” while the ease of use is not satisfied by the user behavior is “Performing” – “Searching” – “Performing”.

**Chapter 4 : Evaluation**

Currently, we can still only perform a rough qualitative analysis of all metric models and can only speculate how high or low each metric is. To obtain exhaustive, systematic, and quantitative models, we need to design experiments to collect eye movement data corresponding to each NFR during real user tests.

## 4.1 Selection of Participants

The participants in the control experiment will be recruited from among people who had some basic knowledge of computer use. Pragmatically, to ensure that the number of participants led to relatively reasonable conclusions, we planned to recruit 20 students who majored in computer science and technology to participate in the experiment.

## 4.2 Experimental Materials and Tasks

The organization of our test material is shown in Figure 7 below. For the hardware device, we used the Tobii Eye Tracker 5 game eye tracker as shown in Figure 4 and Figure 5 above. Using a driver developed based on Streaming Engine, we obtained the raw eye-tracking data from the hardware device and submitted it to the backend application. We designed software under test to simulate a real system with which the user interacts directly. Some code is fed into the program under test to send real-time execution of subtasks and component position to the back-end application.



Figure 5 Tobii Eye Tracker 5 gaming eye tracker Equipment



Figure 6 Actual experimental scenario of eye tracker use

The backend application is divided into two parts, where the HTTP Server accepts and records data from the software under test and the eye-tracking components. The analysis system analyzes the recorded eye-tracking data, subtask execution, and component coordinate positions, including but not limited to extracting high-level metrics, eye-tracking patterns, etc. from the raw eye-tracking data. Finally, the analysis system will report on the NFR satisfaction situation of each subtask.

图示

描述已自动生成

Figure 7 Experimental material organization

The software under test communicates with the data capture platform via a backend application that records the raw eye-tracking data acquired by the hardware device and the execution of subtasks on the front-end and the real-time position of the components, thus providing the basis for later analysis.

We designed a simulated university library system based on a web application as the software under test. The software itself has functions such as information query and seat reservation, a preview of all ages is shown in Figure 8 to Figure 15. We developed the ability to insert intentionally bad designs on it in real-time. We plan to use the application as the target software for evaluating the six typical NFRs described above. We will let participants use the software and focus on evaluating the library seat reservation function.

We will divide the entire process into five subtasks, namely “Navigate to the booking function page”, “Choose the study room venue”, “Choose the date and time”, “Choose the seat” and “Submit booking information”. The specific subtasks and how we inserted intentionally poor designs on them are shown in Table 1below. As participants completed each subtask, we would use the eye-movement data collection system described above to collect their eye-movement data.

Table 1 Experimental grouping and correspondence between intentionally bad design, non-functional requirements, and subtasks

|  |  |  |  |
| --- | --- | --- | --- |
| Test groups | Subtasks | Tested NFR | Intentional design flaws |
| Test group 1 | Navigate to the booking function page | Accessibility | Set the language of the navigation home page to Russian |
| Choose the date and time | Ease of use | The determination range of the selection time is tiny and requires a highly accurate hit on the center of the button |
| Choose the seat | Robustness | Not selecting or selecting more than one seat, the system will crash, forcing the user to re-select |
| Test group 2 | Choose the study room venue | Accuracy | Clicking on the Main Library button will select Library 1; clicking on Library 2 will only select Main Library |
| Choose the date and time | Consistency | The choice of start time and end time provides different design patterns |
| Submit booking information | Device efficiency | After clicking submit button, the page goes to load, artificially reducing the response speed |
| Control group | No intentional design flaws | | |

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

Figure 8 Preview of the home page without inserting intentionally bad design

The subtask “Navigate to the booking function page” takes place on the home page that appears when the user first enters the system, as shown in the figure above. The user needs to recognize and understand the sidebar and top navigation bar, and click on the correct column to enter the seat “booking” function page. Thus, completing the current “Navigate to the booking function page” to complete the current subtask of “Navigate to the booking function page” and start the next subtask.

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

Figure 9 Preview of the home page after inserting the intentionally bad design for accessibility

The non-functional requirement that we inserted with this scenario is accessibility. It is defined as “Degree to which a service can be used by people with distinctive characteristics and capabilities to achieve a specified goal in a context of use”. The deliberate reverse design as: for participants with “no Russian reading and comprehension skills”. The deliberately replaced the guide with Russian to prevent them from using the service and achieving the task goal of the current subtask. The need to improve “Accessibility” of the current page emerged in the course of the participants' use.

图形用户界面, 应用程序

描述已自动生成

Figure 10 Preview of choosing the study room venue page

The scenario of the subtask “Choose the study room venue” is shown in the figure above, which occurs after the user completes the subtask “Navigate to the booking function page”. The first page appears after entering the “Booking” function page. The user needs to select and click on the venue of his choice by recognizing and understanding the guidance in the center of the page. The venue selected by the user will be shaded precisely as feedback to indicate that the venue is selected. The user then needs to click on the “Next” button to complete the current “Choose the study room venue” subtask and start the next subtask.

The non-functional requirement that we inserted in this scenario is accuracy, defined as “Degree of precision to which service provides right”. In the current scenario, it is not the venue that the user clicks on that is selected, but the next venue that the user clicks on that is selected. The need to improve “Accuracy” of the current page emerged in the course of the participants' use.

日历

描述已自动生成

Figure 11 Preview of choosing the date and time page without inserting intentionally bad design

The scenario of the subtask “Choose the date and time”, is shown in the figure above. It occurs after the user completes the “Navigate to the booking function page” and after completing the “Navigate to the booking function page” and “Choose the study room venue” subtask, which is the second page that appears after entering the “Booking” function page. The user needs to recognize and understand the role of the functional components in the center of the page, select the date in the calendar page on the left, and select the start and end time in the two-time selection lists on the right. Both time selection lists use the same design pattern and have a wide range of click judgments for both date and time. The user then clicks on the “Next” button to complete the current “Choose the date and time” subtask and start the next subtask.

The non-functional requirements we inserted in this scenario are “Ease of use” and “Consistency”. We first describe our experimental design for Ease of use, which is defined as “Degree to which a service is easy for users to operate and control”. We made a deliberate reverse design: to make the service difficult to operate and control, in the current scenario, we deliberately narrowed down the range of judgments for the date and time selection controls by changing the range of judgments to one-tenth of the original range. In the current scenario, the user has to be cautious in order to complete the task. The need to improve “Ease of use” of the current page emerged in the course of the participants' use.

日历

描述已自动生成

Figure 12 Preview of choosing the date and time page with inserting intentionally bad design for consistency

Next, we describe our experimental design for consistency, as shown in the figure above, for the definition of consistency, i.e., “Degree to which representative services offer the same design pattern”. In the current scenario, two time-selected controls with identical meanings use two different design patterns: an expanded menu and a drop-down menu. Let the need to improve the “consistency” of the current page arise during the use of the participants.

图形用户界面, 文本, 应用程序

描述已自动生成

Figure 13 Preview of submit booking information page

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

Figure 14 Preview of submit booking information page on loading

The scenario of the subtask “Submit booking information” is shown in the figure above and occurs after the user completes the “Navigate to the booking function page”, “Choose the study room venue” and “Choose the date and time” subtasks. Which is the third page that appears after entering the “Booking” function page. The user needs to click the Next button to confirm and submit the booking information selected in the first two subtasks to complete the current “Submit booking information”. After submission, the system will search the database for currently available seats and display them on the next page. Until the system search is completed, the loading icon will be displayed on the page in a loop and will not go to the next page and the next subtask. The non-functional requirement we inserted in this scenario is “Device efficiency”. It is defined as “Degree of promptness of the service in using a certain number of resources”. The deliberate reverse design as: to reduce the timeliness of the service's return results, i.e., to make the back-end take at least 10 seconds to complete the search in the current scenario so that the participants in the process of using the current page appear to improve the Device efficiency needs.

图片包含 图形用户界面

描述已自动生成

Figure Preview of choosing the seat page without inserting intentionally bad design

The scenario of the subtask “Choose the seat” is shown in the figure above and occurs after completing the “Navigate to the booking function page”, “Choose the study room venue”, “Choose the date and time”, and “Submit booking information” subtasks, which is the fourth page that appears after entering the “Booking” function page, it is also the last page in the whole experiment. Users need to find the seat number they need on the page, click the corresponding icon, and finally click the Next button to complete the subtask.

图形用户界面, 应用程序

描述已自动生成

Figure Preview of choosing the seat page with inserting intentionally bad design for robustness

The non-functional requirements that we inserted in this scenario are Robustness. itis defined as a “Degree that represents the ability of the service to act properly even if some of the input parameters are missing or incorrect”. The deliberate reverse design: to make the service easily crash and fail to act properly if the input parameters are missing or incorrect, i.e., in the current scenario, once the In the current scenario, once the user selects more than one seat or no seat, the service will “crash” and the user must reload the page and retry. This makes it possible for participants to use the service so that the need for improved Robustness of the current page arises.

## 4.3 Experimental Design

We divided the participants into two groups, with ten people in each group. Based on the self-developed test software for inserting bad designs described above, we designed test tasks for different NFRs in different subtasks covering the six typical NFRs we mentioned. In the test for the experimental group, we will turn on intentional bad designs for the participants to trigger their NFRs. These designs are intentionally staggered so as not to interfere with each other and give the participants time to recover from the previous bad designs caused by confusion and dissatisfaction. While in both groups, subtasks with intentionally bad design turned on will be used as experimental group data. In contrast, subtasks without bad design turned on will be used as control group data, which is assigned to maximize the use of the limited number of participants. The control group data will be used as the baseline data for completing each subtask. We will be able to compare the control group with the experimental group to verify whether our theoretical model above is correct. During the experiment, we ask participants to use the “thinking aloud” method to verify that the intentionally poor design does induce their corresponding NFR.

## 4.4 Results

We compared the differences between the collected experimental data and the control data, analyzing whether the differences between them are consistent with the characteristics of our proposed indicator model. Finally, in conjunction with these data, we hope to improve our model by upgrading the judgment of each metric characteristic from a high or low qualitative to a specific scale. In this way, the model allowed to be more precise in its judgments. The experimental results are as follows.

### 4.4.1 Accuracy

The data comparison between the experimental and control groups in the “Select Space” subtask was calculated as follows.

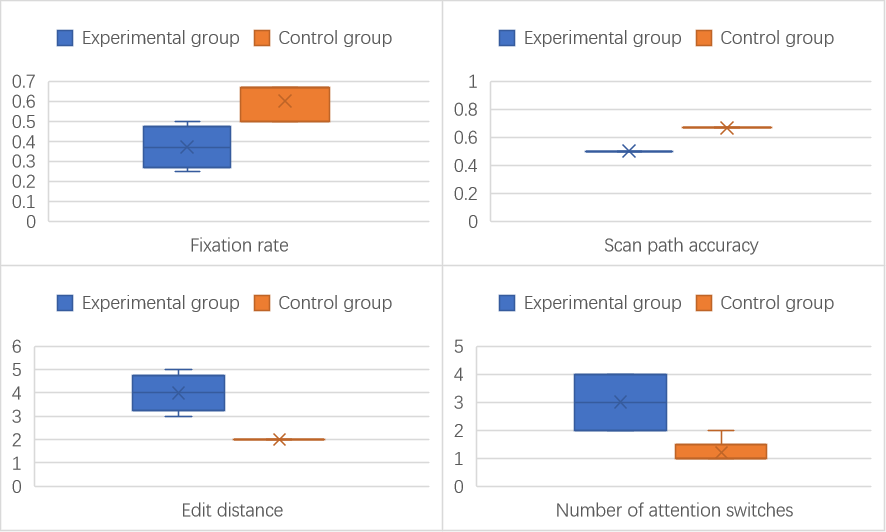


Figure 17 Comparison of four eye-tracking data with significant differences between the experimental and control groups in the “Select Space” subtask

According to the above Figure 17, it can be seen as the analysis in section 3.2. The experimental group’s “fixation rate” is all lower than the control group’s, which can be used as a characteristic. The “Scan path accuracy” because of the small number of AOI areas, and the results are relatively single, but still can clearly show the characteristics of the experimental group average lower than the control group. The experimental group’s “edit distance” is all higher than the control group, which can be used as a characteristic. The “attention switch” of the experimental group is higher than that of the control group, which can be used as a characteristic.



Figure Comparison of four eye-tracking data with no significant differences between the experimental and control groups in the “Select Space” subtask

Figure 18 shows, as analyzed in section 3.2. The “fixation count” of the experimental group contains all the data of the control group, which means that there is no significant difference between them. The “fixation duration” and “first fixation time on AOI” of the experimental group was very similar to that of the control group in both median and mean values, indicating that there was no significant difference between them. The “convex hull area” of the control group contains all the data of the experimental group, which means that there is no significant difference between them. The above performance shows that the performance of these metrics does not change when “Accuracy” is satisfied and not satisfied, so it cannot be used as a characteristic to identify the degree of satisfaction of “Accuracy”.

In summary, when “Accuracy” is not satisfied, it is characterized by low fixation count, low scan path accuracy, high edit distance, and high attention switch. therefore, each parameter of the evaluation formula of the unsatisfied degree is brought into the true value to find

The “Accuracy” dissatisfaction scores for the control and experimental groups are shown below.

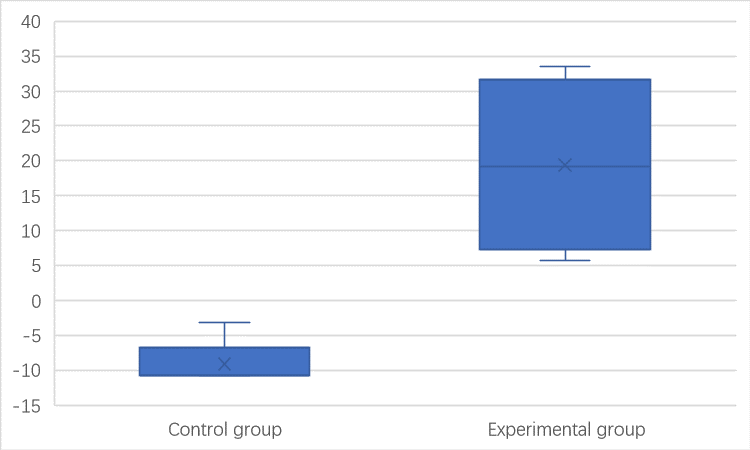


Figure 19 Comparison of “Accuracy” unsatisfied scores in the “Select Space” subtask between the experimental and control groups

This shows that the degree of “Accuracy” unsatisfied formula degree can be bounded by 0 to evaluate whether accuracy is satisfied or not.

### 4.4.2 Robustness

The data for the experimental and control groups in the “Select Seat” subtask were compared as follows.

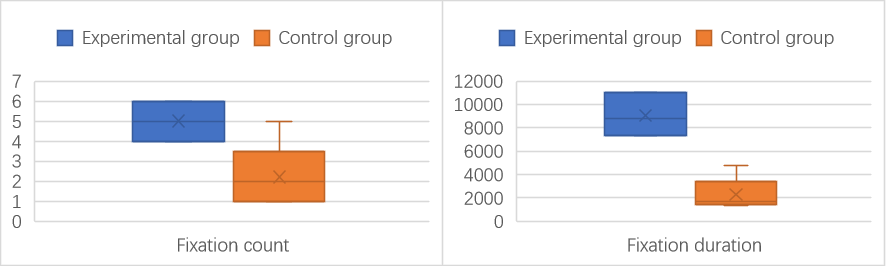


Figure 20 Comparison of two eye-tracking data with significant differences between the experimental and control groups in the “Select Seat” subtask

In the Figure 20 above, it can be seen as the analysis in section 3.2. The “fixation count” of the experimental group is overwhelmingly higher than that of the control group, which can be used as a characteristic. The “fixation duration” of the experimental group is higher than that of the control group, which can be used as a characteristic.

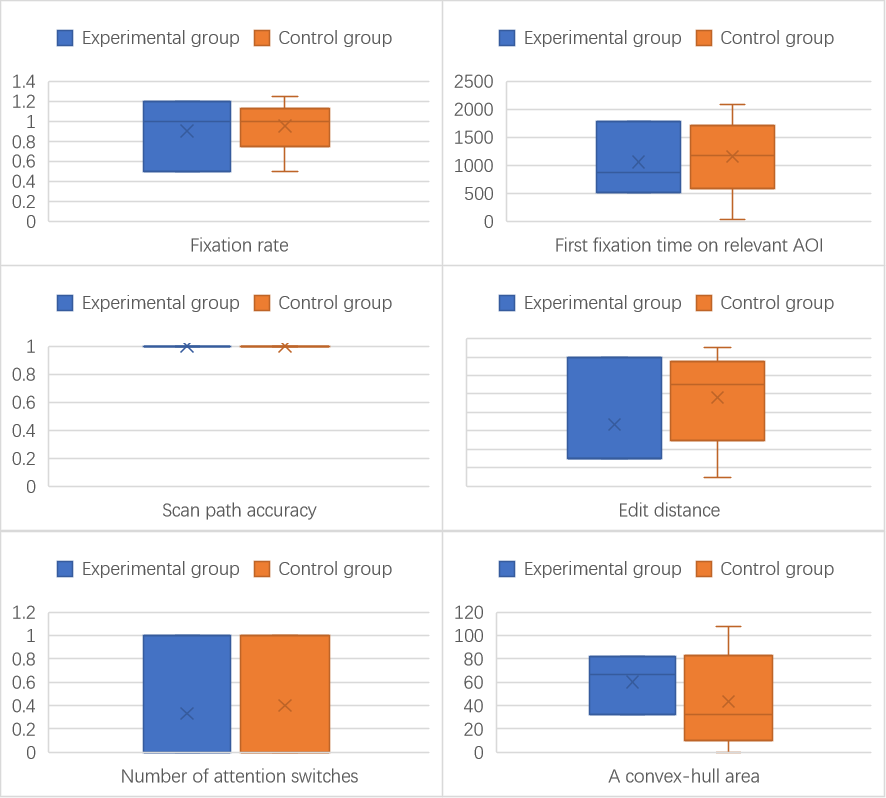


Figure 21 Comparison of six eye-tracking data with no significant differences between the experimental and control groups in the “Select Seat” subtask

Figure 21 shows, as analyzed in section 3.2. The “fixation rate”, “first fixation time on relevant AOI”, “Scan path accuracy”, “Edit distance”, “Number of attention switch” and “a convex-hull area” of the control group contains all the data of the experimental group, which means that there is no significant difference between them. The above performance shows that the performance of these metrics does not change when “Robustness” is satisfied and not satisfied, so it cannot be used as a characteristic to identify the degree of satisfaction of “Robustness”.

In summary, when “Robustness” is not satisfied, it is characterized by high fixation count and duration. Therefore, each parameter of the evaluation formula of the unsatisfied degree is brought into the true value to find:

The control and experimental groups scored “Robustness” dissatisfaction as shown below.

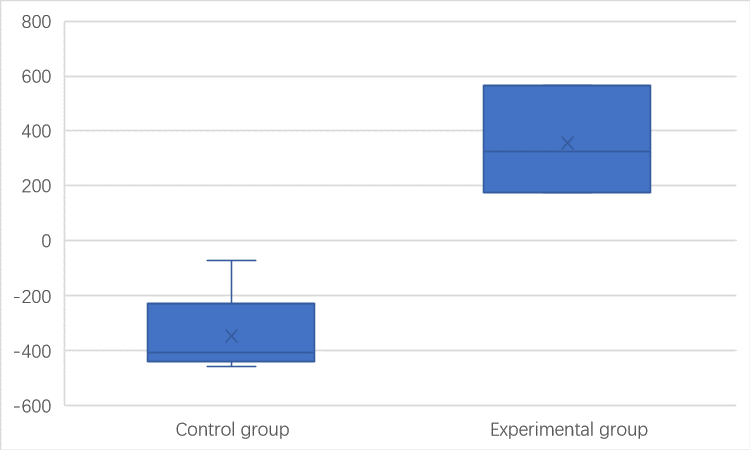


Figure 22 Comparison of “Robustness” unsatisfied scores in the “Select Seat” subtask between the experimental and control groups

This shows that the degree of “Robustness” unsatisfied formula degree can be bounded by 0 to evaluate whether Robustness is satisfied or not.

### 4.4.3 Consistency

The data for the experimental and control groups in the “Choose the date and time” subtask was compared as follows.

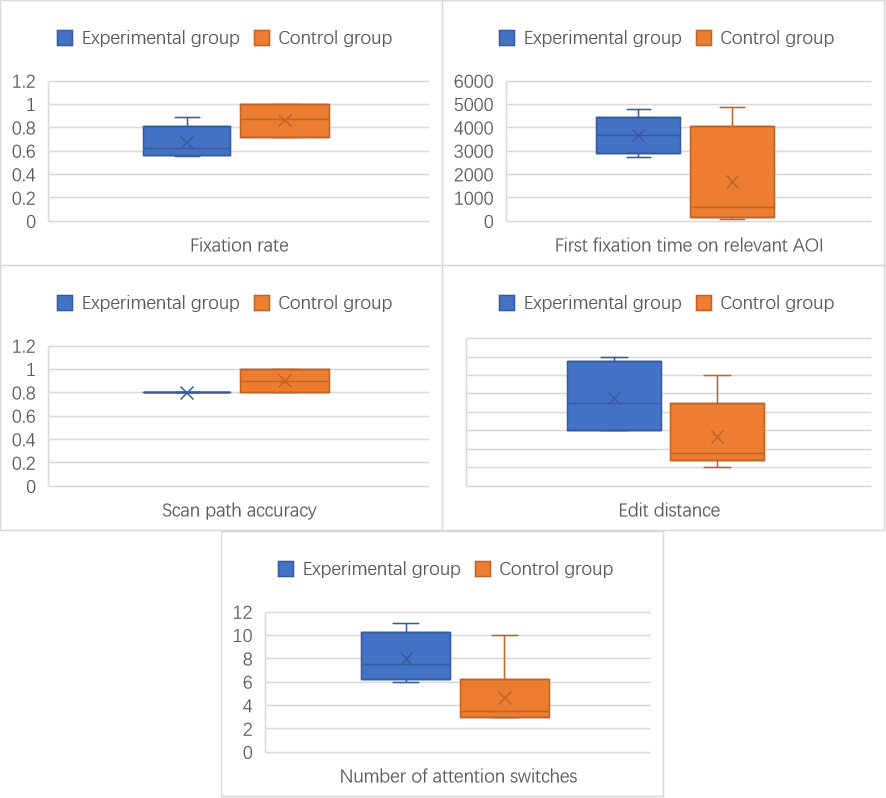


Figure 23 Comparison of five eye-tracking data with significant differences between the experimental and control groups in the “Choose the date and time” subtask

The Figure 23 above can be seen as the analysis in section 3.2. The experimental group's mean and median “fixation rate” are lower than that of the control group, which can be used as a characteristic. The mean and median values of the “first fixation time on relevant AOI ” in the experimental group are higher than those in the control group, which can be used as a characteristic. The “scan path accuracy” in the experimental group was overwhelmingly lower than that in the control group, which could be used as a feature. The mean and median of “edit distance” in the experimental group were higher than those in the control group. The “number of attention switches” in the experimental group was overwhelmingly more heightened than that in the control group, which can be used as a characteristic.

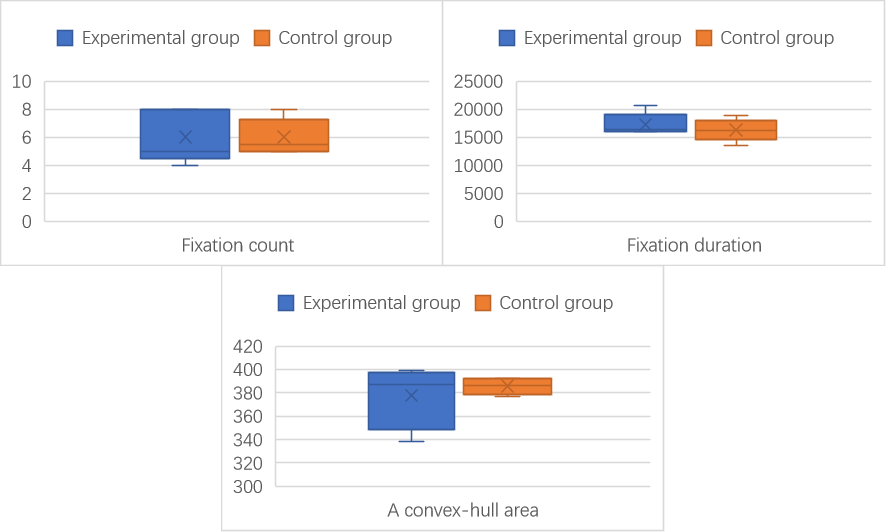


Figure 24 Comparison of three eye-tracking data with no significant differences between the experimental and control groups in the “Choose the date and time” subtask

Figure 24 shows, as analyzed in section 3.2. The “fixation count” and “fixation duration” of the experimental group was very similar to that of the control group in both median and mean values, indicating that there was no significant difference between them. The “a convex-hull area” of the experimental group contains all the data of the control group, which means that there is no significant difference between them. The above performance shows that the performance of these metrics does not change when “Consistency” is satisfied and not satisfied, so it cannot be used as a characteristic to identify the degree of satisfaction of “Consistency”.

In summary: when “Consistency” is not satisfied, it is characterized by a low fixation rate, high first fixation time on AOI, low scan path accuracy, high edit distance, and a high number of attention switches. Therefore, the unsatisfied degree of each parameter of the evaluation equation is brought to the true value to find:

The “Consistency” dissatisfaction scores for the control and experimental groups are shown below.

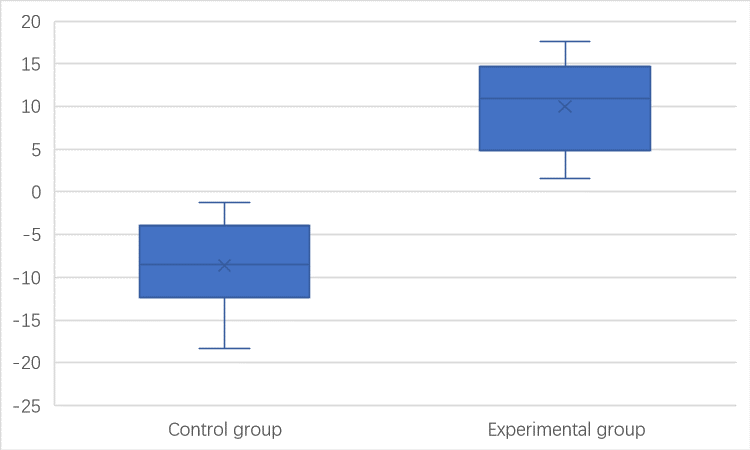


Figure 25 Comparison of “Consistency” unsatisfied scores in the “Choose the date and time” subtask between the experimental and control groups

It can be seen that the “Consistency” unsatisfied degree formula degree can be bounded by 0 to evaluate whether the Consistency is satisfied.

### 4.4.4 Ease of use

The data for the experimental and control groups in the “Choose the date and time” subtask was compared as follows.

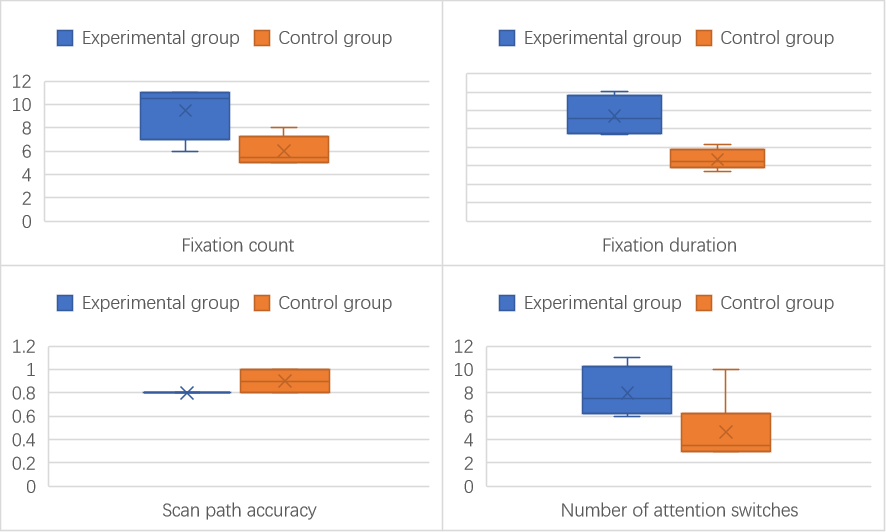
****

Figure 26 Comparison of four eye-tracking data with significant differences between the experimental and control groups in the “Choose the date and time” subtask

According to the above Figure 26, it can be seen as the analysis in section 3.2. The “fixation count” of the experimental group is overwhelmingly higher than the average value of the control group, which can be taken as a characteristic. The “fixation duration” of the experimental group is significantly higher than the average value of the control group, which can be used as a distinct feature. The “scanning path accuracy” is relatively single because of the small number of AOI areas, but it can still clearly show that the average scanning accuracy of the experimental group is lower than the path of the control group, which can be used as a characteristic. The “number of attention switches” in the experimental group is significantly greater than the mean value of the control group, which can be used as a characteristic.

****

Figure 27 Comparison of four eye-tracking data with no significant differences between the experimental and control groups in the “Choose the date and time” subtask

Figure 27 shows, as analyzed in section 3.2. The “fixation rate” of the experimental group was very similar to that of the control group in both median and mean values, indicating that there was no significant difference between them. The “first fixation time on relevant AOI” and “a convex-hull area” of the control group contains all the data of the experimental group, which means that there is no significant difference between them. The “edit distance” of the experimental group was very similar to that of the control group in median values, indicating that there was no significant difference between them. The above performance shows that the performance of these metrics does not change when “Ease of use” is satisfied and not satisfied, so it cannot be used as a characteristic to identify the degree of satisfaction of “Ease of use”.

In summary: When “Ease of use” is not satisfied, it is characterized by high fixation count, high fixation duration, low scan path accuracy, and high attention switch. therefore, each parameter of the evaluation formula of the unsatisfied degree is brought into the true value to find.

The “Ease of use” dissatisfaction scores for the control and experimental groups are shown below.

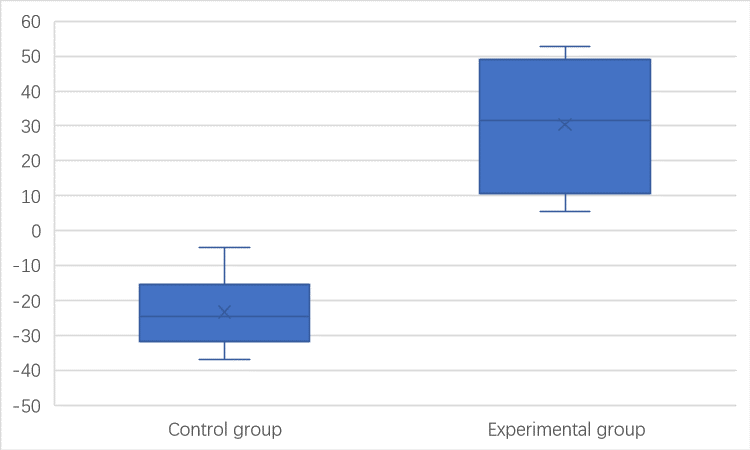


Figure 28 Comparison of “Ease of use” unsatisfied scores in the “Choose the date and time” subtask between the experimental and control groups

It can be seen that the degree of “Ease of use” unsatisfied formula degree can be bounded by 0 to evaluate whether Ease of use is satisfied or not.

### 4.4.5 Device efficiency

Data comparing the experimental and control groups in the “Submit booking information” subtask were counted as follows.

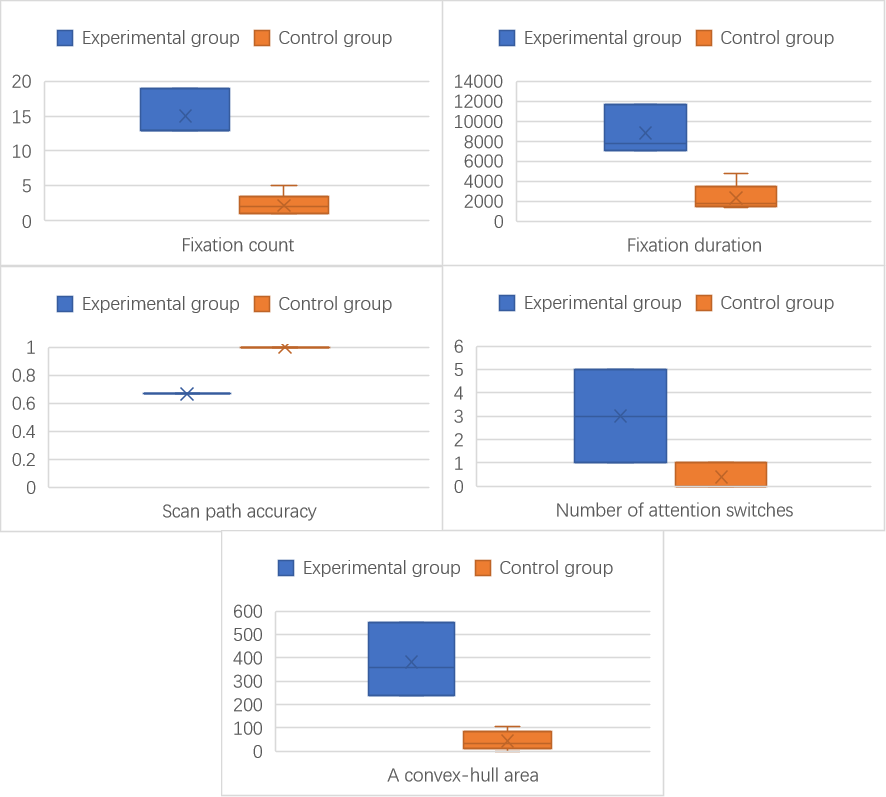


Figure 29 Comparison of five eye-tracking data with significant differences between the experimental and control groups in the “Submit booking information” subtask

The above Figure 29 can be seen as the analysis in section 3.2. The Fixation count of the experimental group is higher than that of the control group, which can be used as a characteristic. The fixation duration of the experimental group is higher than that of the control group, which can be used as a characteristic. The scan path accuracy is relatively single because of the small number of AOI areas, but it can still clearly show that the mean scan accuracy of the experimental group is lower than that of the control group, which can be used as a characteristic. The experimental group’s attention switch is overwhelmingly higher than the control group, which can be used as a characteristic. The convex hull of the experimental group is higher than that of the control group, which can be used as a characteristic.

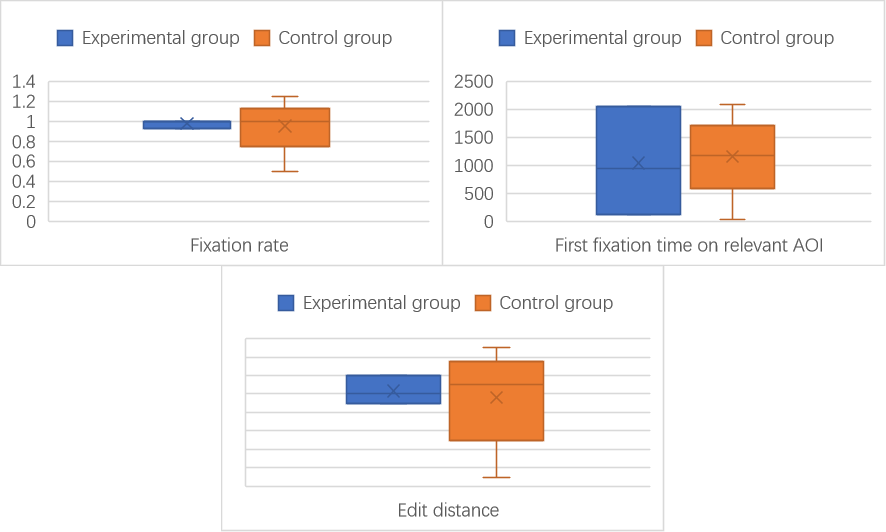


Figure 30 Comparison of three eye-tracking data with no significant differences between the experimental and control groups in the “Submit booking information” subtask

Figure 30 shows, as analyzed in section 3.2. All three metrics “fixation rate”, “first fixation time on relevant AOI”, “edit distance” show that the data of the control group completely contains all the data of the experimental group, indicating that there is no significant difference between the control group and experimental group. The above performance shows that the performance of these metrics does not change when “Device efficiency” is satisfied and not satisfied, so it cannot be used as a characteristic to identify the degree of satisfaction of “Device efficiency”.

In summary: When “Device efficiency” is not satisfied, it is characterized by high fixation count, high fixation duration, low scan path accuracy, high attention switch, and high convex-hull area. therefore, the evaluation formula of the unsatisfied degree of each parameter is brought to the true value to find.

The control and experimental groups scored “Device efficiency” dissatisfaction, as shown below.

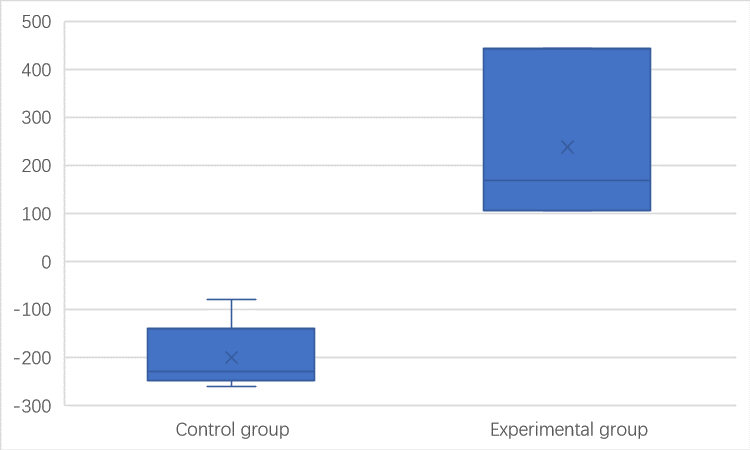


Figure 31 Comparison of “Device efficiency” unsatisfied scores in the “Submit booking information” subtask between the experimental and control groups

It can be seen that the degree of unsatisfied formula degree can be bounded by 0 to evaluate whether “Device efficiency” is satisfied or not.

### 4.4.6 Accessibility

The data for the experimental and control groups in the “Select Seat” subtask were compared as follows.

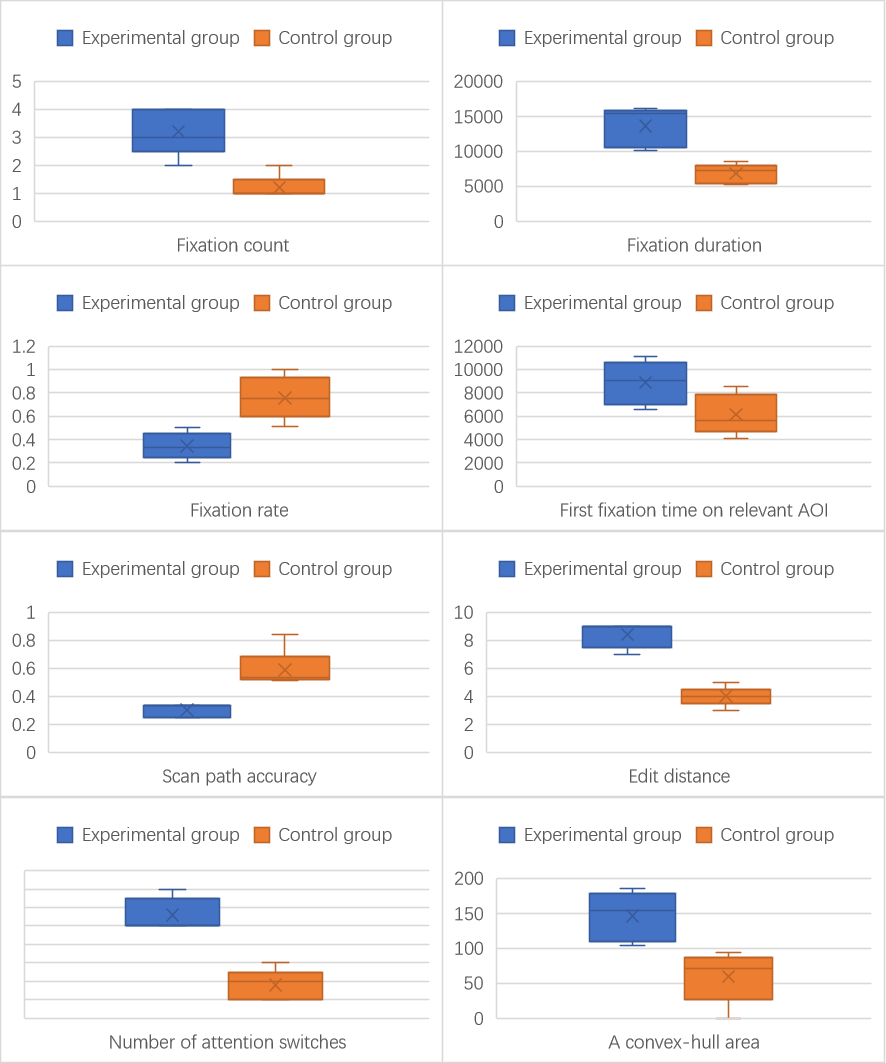


Figure 32 Comparison of eight eye-tracking data with significant differences between the experimental and control groups in the “Select Seat” subtask

In the Figure 32, it can be seen as the analysis in section 3.2. The Fixation count of the experimental group is higher than that of the control group, which can be used as a characteristic. The fixation duration of the experimental group is higher than that of the control group, which can be considered a characteristic. The fixation rate of the experimental group was overwhelmingly lower than that of the control group, which can be characterized. The mean and median of the first fixation on AOI in the experimental group were higher than those in the control group, which could be used as a feature. The scan path accuracy of the experimental group was lower than that of the control group, which could be used as a feature. The edit distance of the experimental group was higher than that of the control group, which could be used as a feature. The attention switch of the experimental group was higher than that of the control group, which could be used as a feature. The convex hull of the experimental group is higher than that of the control group, which can be used as a characteristic.

In summary: When “Accessibility” was not satisfied, the features were high fixation count, high fixation duration, low fixation rate, high first fixation time on AOI, low scan path accuracy, high edit distance, and high; therefore, each parameter of the evaluation formula of the unsatisfied degree is brought to the true value to find.

The “Accessibility” dissatisfaction scores for the control and experimental groups are shown below.

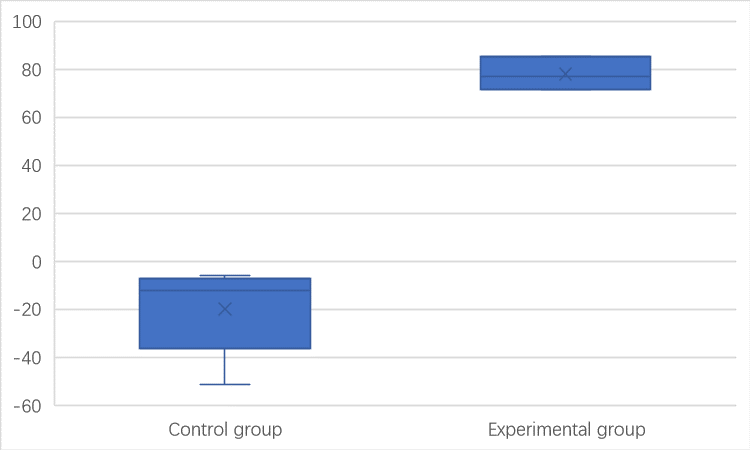


Figure 33 Comparison of “Accessibility” unsatisfied scores in the “Select Seat” subtask between the experimental and control groups

This shows that the degree of unsatisfied formula degree can be bounded by 0 to evaluate whether “Accessibility” is satisfied or not.

## 4.5 Real-World Test

To verify the usability of the whole Eye-tracking Based Requirements Assessment Framework, the non-functional requirements evaluation formulas above-mentioned and have already been parameterized were included. The back-end analysis system can be achieved. For this purpose, two additional participants were invited and asked to re-run the experiments performed in Groups 1 and 2. The difference is that instead of just recording the eye-tracking parameters, the analysis system will directly give them the satisfaction of the nonfunctional requirements in each subtask, with the following results.

Table 2 Real-world test results

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Subtasks** | **Accuracy score** | **Robustness score** | **Consistency score** | **Ease of use score** | **Device efficiency score** | **Accessibility score** |
| **Participant 1** | Navigate to the booking function page | -5.45 | -344.11 | -8.16 | -20.30 | -187.19 | 85.93 |
| Choose the study room venue | -11.91 | -387.69 | -8.76 | -19.93 | -232.94 | -33.91 |
| Choose the date and time | -7.73 | -272.34 | -11.68 | 25.46 | -241.89 | -21.70 |
| Submit booking information | -10.11 | -222.22 | -11.16 | -23.98 | -191.57 | -16.25 |
| Choose the seat | -13.39 | 569.24 | -12.98 | -30.88 | -227.78 | -35.45 |
| **Participant 2** | Navigate to the booking function page | -10.31 | -307.95 | -5.33 | -24.81 | -195.94 | -16.53 |
| Choose the study room venue | 17.93 | -293.57 | -9.88 | -31.43 | -242.98 | -20.35 |
| Choose the date and time | -12.34 | -257.58 | 8.94 | -17.94 | -226.52 | -35.32 |
| Submit booking information | -9.18 | -239.60 | -12.35 | -21.59 | 365.21 | -32.22 |
| Choose the seat | -6.25 | -339.58 | -8.21 | -22.42 | -199.99 | -15.52 |

As shown in Table 2, in subtasks with intentionally bad design for non-functional requirements turned on, the Eye-tracking Based Requirements Assessment Framework can generate correct output for non-functional requirements that are not met, while other non-functional requirements are not misreported. It can be seen that Eye-tracking Based Requirements Assessment Framework can detect the degree of non-functional requirements satisfaction in a real-time environment.

**Chapter 5 : Conclusions and Future Work**

The search for ways to extract valuable information on non-functional requirements from users is paramount and still of interest to researchers. A non-functional property is a property of a whole system and cannot be verified for an individual component. Detection and modification of these functions in the subsequent system maintenance cause high costs. There is no clear and comprehensive solution to assess how non-functional requirements are met.

This thesis presents eye-movement metrics that can effectively distinguish non-functional requirements and use several examples of non-functional requirements to demonstrate how the system works. Furthermore, the approach present with this thesis focuses on providing a way of direct observation of the user that can be used to describe and inform the study of non-functional requirements. This thesis presents the starting point for an integrated and structured approach. The thesis investigates a new method of non-functional requirement analysis, which explores eye movement patterns to collect users' implicit feedback automatically. The thesis established a semantic connection between eye movements and user perceptions based on cognitive and empirical evidence from existing studies. It further formalizes several eye movement patterns and provides a correspondence between them and non-functional requirements. Additionally, it summarizes and quantifies six standard non-functional requirements of evaluation criteria based on experimental data.

The proposed approach provides the details of the precise location implementation that do not satisfy the non-functional requirements and reduces the cost of fixes and enhancements to the system.

Despite the effective approach presented in this thesis, the number of experimental participants recruited for this study is small due to the COVID-19 epidemic. The findings obtained in the study, as the first work combining eye-tracking with demand analysis, are sufficient to serve as inspiration for subsequent work, but the small sample makes it difficult to use the results as generalized conclusions for direct use in a wide range of real-world work. In future work, we hope to recruit more participants so that the conclusions can be more robustly validated and more non-functional requirements can be analyzed.

In future work, we will give methods to identify quantitative analysis for each dimension of eye-tracking metrics to be compatible with more non-functional requirement classifications. We also intend to validate the proposed techniques in the open-source community, combining real projects with detection systems and experimentation to identify potential oversight issues and improvements to our approach. This thesis is the starting point for an integrated and structured approach. The long-term vision is to build a systematic, highly compliant, and customizable rating system that, combined with appropriate development techniques, can facilitate small and medium developers and provide new inspiration for more prominent developers.

In addition, NFRs may be related to each other. This also means that in real-world applications they may affect each other to the extent that they do not receive separate feedback from the user. This is an objective problem, and there are many studies that focus on the interactions between NFRs, but it is not within the scope of our study. In this work, we try to construct the most obvious and simple requirement scenarios that independently link metrics and NFRs according to the definition of NFRs. In our subsequent work, we will try to consider the relationship between NFRs and refine the mapping by combining the results of research on related topics.

For more application scenarios. This project uses a hardware device that is essentially just a camera sensor with an IR fill light to obtain all the required parameters. This means that the framework can be deployed on a large number of devices (PCs with Windows Hello face authentication, Apple mobile devices with Face ID and Android mobile devices with face authentication). With full adherence to the principle of openness and voluntariness, and with proper handling of privacy regulations, we can look forward to an approach that will get feedback from numerous remote, unknown users implicitly. This will provide a new possibility for feedback on user requirements for software systems.

**Bibliography**

[1] SOMMERVILLE I, SAWYER P. Requirements Engineering: A Good Practice Guide[M]. 1st 版. USA: John Wiley & Sons, Inc., 1997.

[2] MAIRIZA D, ZOWGHI D, NURMULIANI N. An investigation into the notion of non-functional requirements[C]//Proceedings of the 2010 ACM Symposium on Applied Computing. New York, NY, USA: Association for Computing Machinery, 2010: 311-317[2022-05-05]. https://doi.org/10.1145/1774088.1774153. DOI:10.1145/1774088.1774153.

[3] CHUNG L, LEITE J C S do P. On Non-Functional Requirements in Software Engineering[C]. 2009.

[4] MYLOPOULOS J, CHUNG L, NIXON B A. Representing and using nonfunctional requirements: a process-oriented approach[J]. IEEE Transactions on Software Engineering, 1992, 18: 483-497.

[5] MAALEJ W, KURTANOVIĆ Z, NABIL H, 等. On the automatic classification of app reviews[J]. Requirements Engineering, 2016, 21(3): 311-331. DOI:10.1007/s00766-016-0251-9.

[6] GUZMAN E, MAALEJ W. How Do Users Like This Feature? A Fine Grained Sentiment Analysis of App Reviews[C]//2014 IEEE 22nd International Requirements Engineering Conference (RE). 2014: 153-162. DOI:10.1109/RE.2014.6912257.

[7] PANICHELLA S, DI SORBO A, GUZMAN E, 等. How can i improve my app? Classifying user reviews for software maintenance and evolution[C]//2015 IEEE International Conference on Software Maintenance and Evolution (ICSME). 2015: 281-290. DOI:10.1109/ICSM.2015.7332474.

[8] BINKLEY D, DAVIS M H, LAWRIE D, 等. The impact of identifier style on effort and comprehension[J]. Empirical Software Engineering, 2013, 18: 219-276.

[9] SUN P, YANG J, MING H, 等. A Multi-layered Desires Based Framework to Detect Users’ Evolving Non-functional Requirements[C]//2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC): 卷 01. 2018: 28-37. DOI:10.1109/COMPSAC.2018.00013.

[10] XIE H, YANG J, CHANG C K, 等. A statistical analysis approach to predict user’s changing requirements for software service evolution[J]. Journal of Systems and Software, 2017, 132: 147-164. DOI:10.1016/j.jss.2017.06.071.

[11] PAGANO D, MAALEJ W. User feedback in the appstore: An empirical study[C]//2013 21st IEEE International Requirements Engineering Conference (RE). 2013: 125-134. DOI:10.1109/RE.2013.6636712.

[12] WANG T, LIANG P, LU M. What Aspects Do Non-Functional Requirements in App User Reviews Describe? An Exploratory and Comparative Study[C]//2018 25th Asia-Pacific Software Engineering Conference (APSEC). 2018: 494-503. DOI:10.1109/APSEC.2018.00064.

[13] RAYNER K. Eye movements in reading and information processing[J]. Psychological Bulletin, 1978, 85(3): 618-660. DOI:10.1037/0033-2909.85.3.618.

[14] ANDREW T. D. Eye Tracking Methodology[M]. Springer Cham, 2017[2022-06-01]. https://link.springer.com/book/10.1007/978-3-319-57883-5.

[15] Eyetracking: Is It Worth It? :: UXmatters[EB]. [2022-06-01]. https://www.uxmatters.com/mt/archives/2009/10/eyetracking-is-it-worth-it.php/.

[16] A Survey on the Usage of Eye-Tracking in Computer Programming | ACM Computing Surveys[EB]. [2022-06-01]. https://dl.acm.org/doi/abs/10.1145/3145904.

[17] SHARAFI Z, SOH Z, GUÉHÉNEUC Y G. A systematic literature review on the usage of eye-tracking in software engineering[J]. Information and Software Technology, 2015, 67: 79-107. DOI:10.1016/j.infsof.2015.06.008.

[18] JACOB R J K, KARN K S. Commentary on Section 4 - Eye Tracking in Human-Computer Interaction and Usability Research: Ready to Deliver the Promises[M]//HYÖNÄ J, RADACH R, DEUBEL H. The Mind’s Eye. Amsterdam: North-Holland, 2003: 573-605[2022-06-01]. https://www.sciencedirect.com/science/article/pii/B9780444510204500311. DOI:10.1016/B978-044451020-4/50031-1.

[19] KARN K S, ELLIS S, JULIANO C. The hunt for usability: tracking eye movements[C]//CHI ’99 Extended Abstracts on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery, 1999: 173[2022-05-31]. https://doi.org/10.1145/632716.632823. DOI:10.1145/632716.632823.

[20] SHARAFI Z, SHARIF B, GUÉHÉNEUC Y G, 等. A practical guide on conducting eye tracking studies in software engineering[J]. Empirical Software Engineering, 2020, 25: 3128-3174.

[21] GLINZ M. On Non-Functional Requirements[C]//15th IEEE International Requirements Engineering Conference (RE 2007). 2007: 21-26. DOI:10.1109/RE.2007.45.

[22] DOERR J, KERKOW D, KOENIG T, 等. Non-functional requirements in industry - three case studies adopting an experience-based NFR method[C]//13th IEEE International Conference on Requirements Engineering (RE’05). 2005: 373-382. DOI:10.1109/RE.2005.47.

[23] CASTRO C, FANTINATO M. Dictionary of Non-functional Requirements of Business Process and Web Services[M]. Sao Paulo, SP. Brazil.: University of São Paulo, School of Arts, Sciences and Humanities, 2018.

[24] WANG C C, HUNG J C, HUANG C H, 等. Advertising Visual Attention to Facebook Social Network: Evidence from Eye Movements[C]//2018 7th International Congress on Advanced Applied Informatics (IIAI-AAI). 2018: 68-73. DOI:10.1109/IIAI-AAI.2018.00023.

[25] SHARAFI Z, SHAFFER T, SHARIF B, 等. Eye-Tracking Metrics in Software Engineering[C]//2015 Asia-Pacific Software Engineering Conference (APSEC). 2015: 96-103. DOI:10.1109/APSEC.2015.53.

[26] SALVUCCI D D, GOLDBERG J H. Identifying fixations and saccades in eye-tracking protocols[C]//Proceedings of the 2000 symposium on Eye tracking research & applications. New York, NY, USA: Association for Computing Machinery, 2000: 71-78[2022-06-08]. https://doi.org/10.1145/355017.355028. DOI:10.1145/355017.355028.

[27] POOLE A, BALL L J, PHILLIPS P. In Search of Salience: A Response-time and Eye-movement Analysis of Bookmark Recognition[C]//FINCHER S, MARKOPOULOS P, MOORE D, 等. People and Computers XVIII — Design for Life. London: Springer, 2005: 363-378. DOI:10.1007/1-84628-062-1\_23.

[28] BUSJAHN T, BEDNARIK R, BEGEL A, 等. Eye Movements in Code Reading: Relaxing the Linear Order[C]//2015 IEEE 23rd International Conference on Program Comprehension. 2015: 255-265. DOI:10.1109/ICPC.2015.36.

[29] PETRUSEL R, MENDLING J. Eye-Tracking the Factors of Process Model Comprehension Tasks[C]//SALINESI C, NORRIE M C, PASTOR Ó. Advanced Information Systems Engineering. Berlin, Heidelberg: Springer, 2013: 224-239. DOI:10.1007/978-3-642-38709-8\_15.

[30] GOLDBERG J H, KOTVAL X P. Computer interface evaluation using eye movements: methods and constructs[J]. International Journal of Industrial Ergonomics, 1999, 24(6): 631-645. DOI:10.1016/S0169-8141(98)00068-7.

[31] BHATTARAI R, PHOTHISONOTHAI M. Eye-Tracking Based Visualizations and Metrics Analysis for Individual Eye Movement Patterns[C]//2019 16th International Joint Conference on Computer Science and Software Engineering (JCSSE). 2019: 381-384. DOI:10.1109/JCSSE.2019.8864156.

**Acknowledgements**

Throughout the writing of this dissertation I have received a great deal of support and assistance.

I would first like to thank my supervisor, Tong Li, for his great foresight in identifying a gap in the related field in the research direction of this project. During more than a year of research, he used his extensive expertise to dispel all my doubts and gave me valuable help in specifying my research questions and methods. Your insightful feedback has prompted me to improve my thinking and has taken my work to a higher level.

The next person I would like to thank most is our counselor, Dai Lu. She has dedicated herself to solving all our problems in school life with a passionate work ethic. Her presence allows us to focus on our classes and projects without worries. She was not only our academic guide and organizer, but also our companion in life, willing to understand and address any of our ideas. I can't imagine a college life without her, it must be several times tougher than what we are going through now.

Finally, I would like to thank my gracious classmates. As someone who is often careless in life, you have always been unstinting with the information and help you have provided, and communicating with you has always let me know where I am. I am honored to have worked, studied, lived and shared four years with all of you wonderful people.

I would also like to thank Fan Gongxiu Honors College for supporting me throughout my college years. The college has trained my research skills with special courses and plenty of projects. More importantly, the College has given us a broad platform and complete freedom to choose the direction of our interest and has been consistently generous in funding our research. My education at the College has built my confidence to continue my research work afterwards.