Usability Study of an Innovative Application in Public Transport by using Hardware-Based Security Technology

Gertraud Schäfer, Andreas Kreisel, Ulrike Stopka

Technische Universität Dresden, Dresden, Germany
{gertraud.schaefer, andreas.kreisel ulrike.stopka}@tu-dresden.de

Abstract. As part of the OPTIMOS 2.0 funding project of the German Federal Ministry for Economic Affairs and Energy (BMWi), a partner develops the TicketIssuance app for secure hardware-based storage of high-priced tickets. The app has implemented a previously unknown technology using the Secure Element and the NFC interface. It is therefore imperative to investigate the usability of the app for a successful market launch. For this purpose, user tests of a prototype of the app were conducted using the think-aloud method. This study analyses the results of five tasks. Test subjects rate the expected and perceived difficulty level for each task. That forms the basis for identifying improvement strategies. The test subjects' performance, the frequency of errors and problems encountered, and the need for moderator's support form the basis for prioritizing usability items within the tasks. The developed structure to determine the test tasks' prioritization and usability items, layout, navigation, handling, wording, system, and data economy offer improvements to increase usability.

Furthermore, the study investigates the determination of a suitable sample size for usability testing.

Keywords: usability testing, think-aloud method, secure elements in mobile devices, public transport

1 Motivation

The OPTIMOS 2.0 project aims to develop an open ecosystem providing technologies and infrastructure for security-critical services via mobile devices. It focuses on the possibility of secure, standardized hardware-based data storage on the available Secure elements (SE). They are either permanently installed in the mobile device, so-called embedded Secure Elements (eSE), or integrated at the Universal Integrated Circuit Card (UICC). Since a wide variety of mobile devices are offered on the market by different Original Equipment Manufacturers (OEM) and Mobile Network Operators (MNO), integrated hardware components are heterogeneous. OEMs and MNOs, as SE-owners, have to provide access to their SEs. They decide who gets access and provide the corresponding key material storing applets on a SE. Individual service providers need to purchase specific applets in order to gain access to the various SEs. For this purpose, the OPTIMOS 2.0 project develops the Trusted Service Manager (TSM). If OEMs and MNOs join the OPTIMOS2.0-ecosystem, the TSM can manage the access to the different SEs. Service providers get access to the various SEs via the TSM. It acts as the only contractual partner who manages the data or applets stored on all SEs in personal mobile devices.

One main developed application in the project provides the possibility to transfer the personal electronic identity (eID) from the ID card to a personal mobile device securely stored on its SEs. With this technology, it will be possible to enter verified personal data from the derived eID in registration formulas directly transferred without using the physical ID card. Therefore, on the one hand, registration processes can be carried out much faster and more conveniently for customers in a single-step process without manual data entry. On the other hand, service providers using this technology receive verified data minimizing time-consuming verification processes.

Besides, the OPTIMOS 2.0 technology with its central TSM platform offers a wide range of other application scenarios for security-critical applications for mobile services, such as the storage of car keys for car-sharing services, room keys in hotels, or the storage of high-priced tickets in passenger transport.

This study dedicates to the usability of the TicketIssuance app prototype developed for Berlin's public transport operator BVG as part of the OPTIMOS 2.0 project. At the time, the app developers have not yet realized all final application scenarios. One of the primary objectives is to find out how potential users perceive the SE's configuration and the handling of the NFC interface for data transfer to smart cards.

Concerning Berlin's public transportation system, there is a general interest in using apps. In 2018 alone, People downloaded the Fahrinfo app of the Berliner Verkehrsbetriebe (BVG) around four million times [5]. Over a thousand subjects took place in a Germany-wide study surveying the use of public transport apps and electronic ticketing. It shows that about two-thirds of the Berlin study subjects use mobile apps for local public transport, but only 20% had already purchased electronic tickets via apps in 2018. However, 75% of all users see ticket sales as an essential part of public transport apps' functional scope [14]. Therefore, the question is how to design apps so that users increasingly purchase electronic tickets meeting the requirements for fast and secure purchasing and ticket control processes, including convenient payment processing and required data protection to reduce uncertainties and insecurities on the customer side.

New applications need to be developed with a high level of usability to ensure a good user experience. For successful market penetration, developers of innovative applications have to ensure user acceptance, not technical feasibility. Therefore, new apps must be examined about their usability during their development process.

2 Study Object - The BVG TicketIssuance App based on OPTIMOS 2.0 technology

The Berlin public transport company BVG wants to push the distribution of electronic tickets via mobile apps. The currently available BVG apps offer a selection of the most popular tickets in the lower price segment up to a maximum of 84 euros (monthly ticket

Berlin tariff zone AB) by QR code. It is relatively easy to copy. Therefore the BVG faces the risk of service fraud through duplicated tickets. To offer higher-priced tickets, such as annual or monthly once even in the broader tariff zone Berlin ABC, and to use NFC technology advantages for checking tickets, one OPTIMOS 2.0 project partner develops the TicketIssuance app. The app developers are in intense competition with other ones. Developers must be aware of how users can quickly and effortlessly substitute their app due to other ones. The YouGov report 2017 [28] surveyed that 89% of 2000 study subjects had already deleted apps at least once. The most common reason was an uninteresting or disappointing app performance. Therefore, before introducing the TicketIssuance app, it is crucial to ensure usability to achieve user acceptance through a positive user experience and reduce possible uncertainty when purchasing tickets.

As already described, the TicketIssuance app should offer the option to store tickets on the SE in the smartphone. This option requires that the integrated Smartphone SE supports the open OPTIMOS 2.0 interfaces and is addressable via the OPTIMOS 2.0 TSM. In summer 2020, only Samsung's Galaxy smartphones from model S9 and higher were ready to be part of the OPTIMOS 2.0 ecosystem. Other manufacturers do not give access to the SEs on their mobile devices up to now. The same applies to the MNO UICC.

The TicketIssuance app must be able to address the SE. Therefore, the first step is installing the OPTIMOS 2.0 TSM-API app for secure communication between the ticket provider and the SE on the personal smartphone. The one-time initialization process automatically downloads, installs, and personalizes the SE applet as a prerequisite for storing higher-priced tickets in the SE in a forgery- and copy-proof manner.

The checking of tickets stored in the SE occurs via NFC technology and works even on the smartphone's standby or low battery mode. In this way, public transport users can prove their ticket's validity even if the battery level is insufficient. The checking process is similar to one of the tickets stored on smartcards. This technology also allows storing tickets directly to appropriately configured smart cards via the TicketIssuance app. It is also possible to transfer not to copy tickets between SEs of different mobile devices or to a smart card via NFC. This function offers customers the possibility to buy tickets for other persons or to transfer already purchased tickets. One use case could be the purchase of electronic tickets for children or persons using smartcards. The study's prototype app provides low-priced tickets to store as QR code on the internal smartphone storage or a corresponding smart card and higher-priced tickets, such as monthly tickets or the so-called field test ticket, on the smartphone SE or a corresponding smart card.

In its final version, the TicketIssuance app will also provide users with a quick and easy way to personalize a user account by electronic transfer of personal data from the derived digital identity stored at the SE.

By integrating the OPTIMOS 2.0 technology, the final version of this app will offer the following functions:

1. Creation of user accounts with the integration of the eID stored in the smartphone

2. Purchase includes electronic payment and management of different ticket options, exceptionally high-priced season tickets such as monthly and annual tickets, to store securely on the SE.
3. Ticket transfer via the TicketIssuance app to an external smartcard or another smartphone.
4. Checking Ticket stored on the SE via the NFC interface - touching the smartphone to a checking device without opening the app-similarly the smartcards' ticket check; Check is also possible in switched-off mode or when the battery is low.

The usability test study object is the ticket purchase and storage as a central function of the TicketIssuance app prototype. However, subjects have to configure the SE and create and personalize a user account in the first step. The storage options will investigate whether the functions of managing tickets (purchase, show/check, delete) are easy to understand and perform. In particular, the study examines the different possibilities of ticket storage

- Internal at the app of the smartphone,
- At the SE or
- At the smartcard.

3 Usability Testing Methods

Usability testing is "[...] activities that focus on observing users working with a product and performing tasks that are real and meaningful to them" [4]. Depending on the development of an application product, we can distinguish between formative and summative evaluations. The formative evaluation takes place as relatively small studies to identify fundamental usability problems and understand user requirements during the product development process. They provide developers customer-oriented input for the further development process. The focus is on questions such as:

- What works well for the user and what works poorly?
- What are the most critical problems use an application in terms of usability?

After the necessary adjustments to the product, a new evaluation usually occurs. Summative tests take place with a larger sample after completion of the product development. Here it is necessary to ensure the required statistical validity in order to be able to make reliable statements meeting the defined goals, such as a certain degree of efficiency, specific customer expectations, like error frequency, and the time required to fulfill the task or the comparison with competitor products [4, 6, 2]. Usability tests can occur in a laboratory, in a specific room, or under real conditions in the field, such as shopping malls, parks, vehicles, or the customer's home [4].

Various methods are possible to gather information about product characteristics. Card sorting - a participatory design method – is an often applied method in an early stage of product development, collecting user understanding and preference, for instance, developing a user-friendly structure and navigation of a software application [4].

The heuristic evaluation method takes a different approach. A group of usability experts independently tests a software product using defined simple and general criteria (heuristics) to uncover usability problems. The experts evaluated and prioritized their results jointly for processing. [4] One basis is the ten usability heuristics established by Nielsen in 1994 [17].

The think-aloud method can give a good insight into customer experiences in a more advanced application development stage. Test subjects, usually potential users, test the application and think aloud. Barnum summarizes it as follows: "… thinking out loud provides a rich source of information about the user's perceptions of the product's usability" [4]. Nielsen describes it as one of the essential methods for assessing usability [21]. It is also the applied method in this study.

3.1 Think-Aloud Method

Test subjects speak their thoughts aloud while performing the tasks. The think-aloud method follows synchronous verbalization of cognitive processes while performing a specific task performance. Researchers can observe and record mental processes affecting the test during actions or products' use through communication.

Compared to retrospective verbalization, the advantage is mainly the data's consistency and completeness [8]. The data collected are a snapshot regarding the subject- or customer-specific perception using a product or an application [7]. Deep insights into the subjects' problem-solving behavior become apparent. Thus, it becomes apparent which usability problems occur [9]. According to Henry et al. [13], thinking aloud does not affect performance levels. Alhadreti and Mayhew [3] also conclude that findings after analyzing three think-aloud studies.

According to Ericsson and Simon [8], there are different forms of verbalization.

Level 1 verbalizations are expressions as they occur in self-talk. Since they do not require any cognitive processes, there is no significantly higher effort and time requirement than situations without communication. In this way, the thoughts and information articulated are essential in detecting problems and errors in an application. They are expressed spontaneously and are not known to the subject in advance. [8]

Level 2 verbalizations represent an extension of the articulated information from level 1, in that thoughts occurring intuitively in short-term memory are additionally explained and described. In this process, the subject converts thoughts into words. Due to the increased processing time involved, task performance may take longer during the test situation. It does not disturb the general course of the processing and influence the success of the task. It is to assume that the data collected for Level 2 verbalization are reliable. [8]

Level 3 verbalizations reproduce thoughts themselves and explanations concerning the cognitive processes (e.g., behavioral descriptions or motives). Level 3 verbalizations link individual thoughts and memories, which change the process structure and lead to a change in the performance and correctness of the results and an increased time requirement. [8]

Generally, the test design should formulate the think-aloud method's tasks on level 1 and level 2 verbalizations. Accordingly, test subjects should only think aloud and not explain. They should behave as if they were alone in the room. Before starting the test, the moderator should also point out that he will invoke an invitation to think aloud in the phase of prolonged silence. Any distraction should not occur to avoid level 3 verbalizations.

3.2 Determination of the Sample Size for a Usability Test

The question of the required sample size to determine a specific proportion of errors and problems in the context of usability testing is a much-discussed topic in practice and theory. Nielsen/Landauer developed the widely used and simplified model that "[...] is sufficiently accurate for practical purposes" [19]. The formula $Found(\iota) = N (1 - (1 - \lambda)^{\iota})$ $(1Found(\iota) = N (1 - (1 - \lambda)^{\iota})(1)$ calculates the number of errors detected by at least one of the subjects

$$Found(\iota) = N (1 - (1 - \lambda)^{\iota}) \qquad (1)$$

The result depends on the total number of errors N and the proportion of all usability problems a tester will detect (λ) [19]. Thus, researchers have to estimate the number of errors to determine a suitable sample size. Nielsen [20] assumes that one test subject reveals 31% of the errors, and 15 test subjects almost detect all errors. According to this calculation, three to four test subjects can perceive more than half of the errors.

In case the usability test is part of an iterative design process with regular test procedures for the adjustment and advancement of the product, a sample size with five test subjects gives the optimal cost-benefit ratio in practice. This sample size already considers that one or two people will not appear for the test even having agreed to do so. [6, 20]

Virzi [27] supports these assumptions by the finding that four to five test subjects detect about 80 % of the problems. It illustrates that as the number of test subjects increases, the probability of uncovering new information steadily decreases. Furthermore, it assumes that the first test subjects already detect problems that strongly influence the ease of use. The number and level of difficulty of newly uncovered information steadily decrease with increasing sample size.

Usability researchers question and criticize these theories from 1990 to 2000, especially when examining applications with increasing complexity. According to this theory, websites or applications have many functions and operating options, where users have different options at their disposal to achieve their personal goals. It can also influence the proportion of errors (λ) that a tester can uncover. For example, in their study of four different e-commerce websites, Spool/Schroeder [25] could not identify a $\lambda \geq 0.16$ in each case. According to Nielsen's model and this measure, a sample size of at least ten test subjects can detect 80% of the problems and errors. The results correspond with the study of Lewis [16], testing different office systems, according to which the average probability for the uncovering of an error lies with 0.16.

The different findings show there are not any binding statements for an optimal sample size. According to Cockton, this also depends on the following factors [6]:

- "[...] diversity of subjects,
- Test protocol design,

- Variety of task performance,
- Complexity of application,
- Design quality [...],
- Problem reporting procedures,
- Usability goals [...]".

4 Study design

The study design follows the steps visualizes in figure 1.

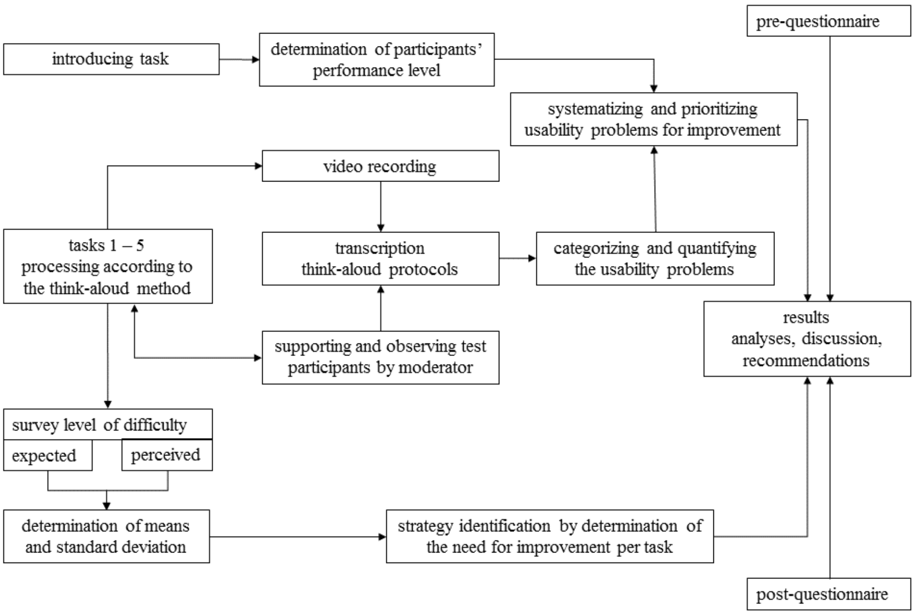
pre-questionnaire
introducing task
determination of participants' performance level
systematizing and prioritizing usability problems for improvement
video recording
tasks 1 – 5 processing according to the think-aloud method
transcription think-aloud protocols
categorizing and quantifying the usability problems
results analyses, discussion, recommendations
supporting and observing test participants by moderator
survey level of difficulty
expected
perceived
determination of means and standard deviation
strategy identification by determination of the need for improvement per task
post-questionnaire

Fig. 1. Study Design.

4.1 Workshop Design for Usability Testing of the TicketIssuancce App

Determination of the evaluation procedure. The TicketIssuance app for BVG to be tested is in the development phase. So far, up to now, no tests with potential users have taken place. This usability study follows the formative evaluation approach.

Determining the Test Environment. A standard method for usability tests following the formative evaluation approach is the laboratory test. For the present study, using a real test laboratory of the BVG is not possible. Nevertheless, the premises at BVG provide a reference to the company and offer a controlled environment for observing the test subjects' behavior during working on the structured tasks. Besides, supplementary data collection methods are used before, during, and after the experiment for extended analysis of user behavior and user requirements.

Determining the sample size. According to formula (1), an exact determination of the influencing variables (especially λ) for calculating the sample size is not feasible. No data is available so far since this is the first usability test. The usability test of the TicketIssuance app should reveal as many problems as possible and provide information for improvement.

The decision is to recruit 20 test subjects. It bases on Nielsens' assumption that a heterogeneous target group, such as public transport services users, will be covered by larger samples [20]. This number also takes into account the non-appearance of invited subjects.

Application of the Think-Aloud Method. The usability test takes place using the think-aloud methodology. The designed test tasks consider an expected processing time and a structure that allows synchronic verbalization (levels 1 and 2). Subjects do not have to resort to long-term memory, which ensures a high level of accuracy. According to Albert and Tullis [2], the following patterns of behavior can occur:

- "[...] Verbal expressions of confusion, frustration, dissatisfaction, pleasure, or surprise.
- Verbal expressions of certainty or indecision about a particular action that may be right or wrong
- Subjects not saying or doing something they should have done or said
- Nonverbal behaviors such as facial expressions or eye movements."

During the test, the moderator's interactions are limited to actions when there are too long pauses in the speech. Then he should ask subjects to speak their thoughts aloud, using short and concise instructions. The moderator documents the frequency and timing of verbal interactions for data analysis. Thus, a repeated request to communicate aloud may indicate a significantly cognitively demanding processing step. Especially concerning later investigations, the experimental structure and execution allow for later comparability and aid in interpretation. [22, 23]

A test design with the think-aloud method shows apparent differences compared to everyday social communication. The test moderator receives instructions in a created moderator script for stringent execution. He explicitly points out that the object of investigation is the app itself and the associated recognition of problems using the app, not the subject's abilities.

Design of questionnaires. A short standardized questionnaire before and after the usability-test of the app collects a few supplementary data.

Pre-questionnaire. Test subjects should get a comfortable and pleasant introduction to the unfamiliar test situation. Therefore, the pre-questionnaire serves to collect the sample's socio-demographic data and data on the previous use of BVG services. Furthermore, these data are the basis for describing the sample description and the result's analyses. Age cohorts base on empirical studies from the field of transportation for comparability [11].

Post-questionnaire. The post-questionnaire collects some supplementary information after the usability test's task processing has been completed. One criterion for evaluating usability is satisfaction. A questionnaire can best capture this [12]. Thus, one question focuses on the satisfaction of the tested app's usability using a five-point scale from one (very satisfied) to five (very dissatisfied). Other questions elicit the importance of using different media for electronic ticket storage and the subject's knowledge about the NFC interface of the personal smartphone.

Test Tasks. The concept of the tasks takes into account that no results from usability tests are available so far. There are no known indications where usability difficulties will exist. Therefore, the test tasks follow real future use cases.

The introducing task determines the test subjects' skills and abilities in dealing with mobile apps. For this task, subjects use their smartphones to determine their performance level. The procedure assumes that individuals with a comparatively low-performance have a less common approach to using apps and uncover a higher number of problems or errors. The basis for determining the performance level is the time required in each case.

In task 1, the test subjects get the instruction to install the app on the test smartphone. To do this, they must open, download and install the TicketIssuance app in the Playstore. Task 2 creates the precondition for ticket purchase and storage. Test subjects have to configure the app by initializing the SE in the first step. It is the prerequisite for the storage of high-priced tickets in the smartphone. In a further step, they create a user account and personalize it via manual data entry. With task 3, the test subjects are to purchase various tickets and store them on different storage media - in the internal database memory, on the SE, or the smartcard. Task 4 requires test subjects to retrieve the purchased and stored tickets, show them for a potential control operation and delete them afterward. In task 5, they shall remove the user account, uninstall the SE and delete the app.

4.2 Survey and Evaluation Methodology

Identification of Strategies by Determination of the Need for Improvement per Task. One usability factor is satisfaction, which can vary widely based on different expectations and perceptions. In determining satisfaction, the confirmation/disconfirmation paradigm helps. It assumes that people compare the perceived performance level after using an application with their expectationss before using it [1]. If the level of perception is above the expectation level (positive disconfirmation), this leads to a positive perception of quality and satisfaction. Dissatisfaction (negative disconfirmation) occurs if the perception level is lower than the expected one. If both levels are equal, there is neither explicit satisfaction nor dissatisfaction. Against this background, the test subjects indicate for each task how difficult they expect this task to be or how difficult they perceived the processing to be. [1, 15].

There are four derived strategies, depending on the average values of expected and perceived level of difficulty. The strategy "No Modification" implies that users expect and perceive simple task handling. Thus, there is, for the time being, no need for modification. They will not result in an increased quality perception or customer satisfaction. The strategy "Improve Immediately" contains tasks for which the test subjects expected significantly easier processing than was possible in the end. There is a high potential for improvement in these tasks. It needs a priority effort to correct these identified errors and problems. If persons expect many difficulties and perceived much less, then these tasks can attract customers and lead to high satisfaction. These tasks belong to the strategy of "Advertise". They are starting points for communication with potential customers, particularly in the market penetration phase, highlighting potential benefits. However, expectations will change with frequent use of the app if users can draw on their experience. For tasks in the fourth strategy "Good Opportunity" the expected and perceived level of difficulty is generally higher than for the other tasks. Improving usability can lead to a more positive perception of quality and increased satisfaction. In this case, users will also adjust their expectations based on experience with the application and similar products. The individual strategies' delimitation uses the median of the values collected, assessing the expected or perceived level of difficulty based on a five-stage scale (1= very simple to 5 = very difficult) [1].

Test subjects have very different abilities and skills in using apps, on the one hand, regarding smartphone use and, on the other hand, how often they use which apps for specific tasks, e.g., ticket purchase, connection search for public transport services. When interpreting the usability test results, the test subjects' performance levels can be significant. This determination of performance levels uses a standardized test based on completing the introducing task. The average of the individual times is the norm time. It is the basis to determine the test subjects' performance levels concerning their respective personal performance for classification and comparison of individual abilities. [26]

Frühwald [10] calculates the degree of performance as follows:

$$performance\ level = \frac{norm\ time}{observed\ individual\ time} \times 100\ [\%] \tag{2}$$

However, this norm-oriented test does not claim to define subjects' performance concerning their total smartphone use and handling ability.

Systematizing and Prioritizing Usability Problems for Improvement. The perceived problems during the test are the basis for recommendations improving the app's usability. One method is a discussion round in the test team, which determines the necessary adjustments' problem prioritization [4]. Even usability experts can classify the identified problems. Both methods have subjective influences depending on the diverse personal experiences, affecting the quality of the results [18].

The present study uses an own developed methodology based on Nielsen [29] to prioritize problems by importance. It considers two dimensions. The first one incorporates the average performance level scores of subjects who perceive the problem and the detected problems' frequency. A strong influence exists if the perceived problem occurs with above-average frequency during task processing or if the subjects' average performance level is low. The second dimension considers the possibility of independent task processing. A strong influence exists if the test subject needs the moderator's support for a task processing successfully.

Table 1 categorizes the collected problems and systematizes them by priority for improvement.

Table 1. Systematization of the surveyed usability problems according to the importance for improvement.

Dimension 2: Influence of problems or errors on the self-reliant task processing	**Dimension 1: Influence of the test subjects' performance level plus the occurred error frequency**	
	High low average performance level of test subjects or high frequency of errors	**Low** high average performance level of test subjects and high frequency of errors
High successful task processing not possible without moderator's support	**Priority 1** Detection of errors very important Need action immediately	**Priority 2** Detection of errors important Need for remedial action before introduction into practice
Low self-reliant successful task processing possible without moderator's support despite usability problem	**Priority 3** Detection of errors preferable	**Priority 4** Detection of errors after fixing the others

The basis for quantification and prioritization are the defined design areas (1) navigation, (2) layout, (3) handling, (4) wording, (5) system, and (6) data economy, to which the specific problems are assigned.

Correlation between number of errors and number of interactions. In the context of usability testing by the think-aloud method, the test moderator should interact with the test subjects as little as possible. High usability does not need any intervention of the moderator. However, test users should detect as many usability problems as possible. The tasks follow a workflow of real use. Therefore the test subjects have to finish each. This goal can sometimes require the intervention of the moderator. There is an expectation of a positive correlation between the number of problems and the number of necessary moderator interactions related to the individual tasks.

This analysis uses the empirical correlation coefficient according to Bravis and Pearson as a measure of linear correlation as well as the rank correlation coefficient of Spearman, which measures the monotonic correlation of two variables [24]. The correlation coefficients will help the result's analysis and interpretation.

5 Study Results

5.1 Sample Description

Of the 20 selected and invited test subjects, eleven appear. That corresponds to the minimum number of test subjects for uncovering more than 80% of the problems and errors. The group is sufficiently homogeneous. At least one subject from each of the four age cohorts participates. Most subjects belong to cohorts between 25 and 64 years. All test subjects are regular users of public transport services and use different types of tickets. Thus, they could draw on very different experiences and have different expectations regarding the purchase of tickets. The same applies to the experience of using apps for public transport. Almost all subjects know the four different BVG apps. However, only a few use them frequently or occasionally.

5.2 Analysis of the test task

The improvement of usability serves to achieve broad user acceptance and high user satisfaction. The expected and perceived levels of difficulty of the five test tasks form the basis for positioning them in the developed matrix. The subdivision of the two coordinates bases on the median by all 55 ratings from the eleven subjects. The respective average mean values determine the positioning of the tasks into the strategy fields (cf. Table 2, Fig. 2).

Table 2. Expected and perceived level of difficulty per task.

	Level of Difficulty (1= very simple …. 5 very difficult)			
	Average mean		Standard deviation	
	Expected	Perceived	Expected	Perceived
Task 1	1,18	1,27	0,4045	0,6467
Task 2	1,73	2,36	0,6467	0,9244
Task 3	2,77	2,27	0,9840	0,7862
Task 4	1,55	2,27	0,6876	0,6467
Task 5	2,00	2,00	0,7746	0,8944

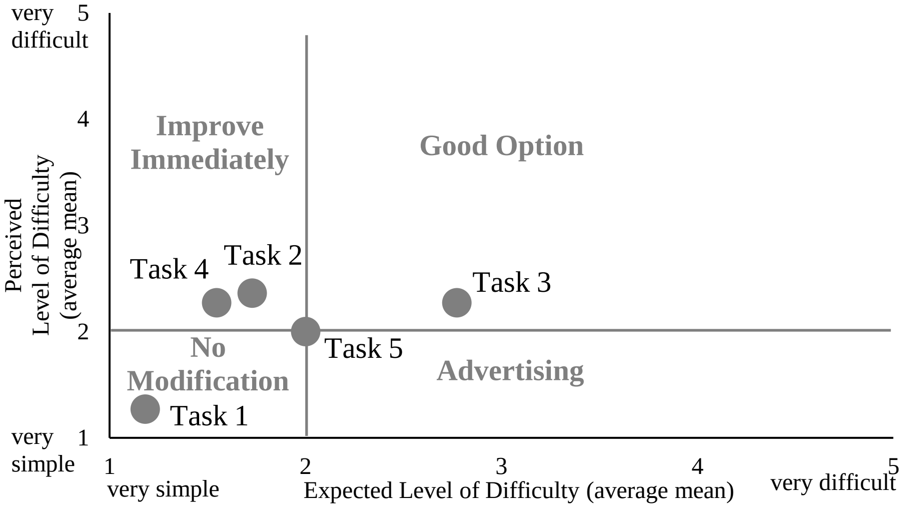

Fig. 2. Strategy identification by Determination of the need for improvement per task.

Task 1. The test subjects do not have any particular difficulties with the installation of the app. Therefore, the task assignment falls to the strategy "No modification" (cf. Fig. 2). The mean values of the expected and perceived difficulty levels are the smallest compared to the other tasks. The likewise low standard deviations show a relatively homogeneous evaluation. In total, there are seven interactions between the test moderator and subjects. The test moderator intervenes once directly to ensure successful task completion. The tests identify one problem in the fields of action layout and wording (cf. Table 3). For test task 1, in general, there is no prioritized need for action.

Table 3. Task 1 - detected usability problems.

Dimension 2: Influence of problems or errors on the self-reliant task processing	**Dimension 1: Influence of the test subjects' performance level plus the occurred error frequency**			
	High		**Low**	
	Field of action	Problems	Field of action	Problems
High	**Priority 1**		**Priority 2**	
			Layout	1
Low	**Priority 3**		**Priority 4**	
			Wording	1

Task 2. Test task two involves processing steps to create the prerequisites for app use. That includes setting up and personalizing a user account and configuring the SE for storing high-priced tickets.

The observed usage behavior is quite different from task 1. The tests detect a total of 21 usability problems in all six fields of action. In total, there are 58 interactions between the test moderator and subjects. Direct intervention by the test moderator for successful task completion is necessary in 13 cases.

These results also reflect the ratings on the expected and perceived difficulty level. That leads to the positioning of the task in the matrix field "Improve immediately" (cf. Fig. 2). The test subjects expect the task processing to be significantly easier before the test than they perceive them during the test. The expected difficulty level's measurement scatters less around the mean with 0.65 than the perceived one's values with 0.92. The expectation is more homogeneous than the perception. The processing steps to be performed in task two involve the SE configuration, a process unknown to the test subjects. They have no previous experience or operating analogies to fall back on. However, for the app's full use, the users must configure the SE and set up a user account. The fixing of discovered usability problems should be prioritized according to Table 4 and corrected as quickly as possible to avoid acceptance barriers.

Table 4. Task 2 - detected usability problems.

Dimension 2: Influence of problems or errors on the self-reliant task processing	**Dimension 1: Influence of the test subjects' performance level plus the occurred error frequency**			
	High		**Low**	
	Field of action	Problems	Field of action	Problems
High	**Priority 1**		**Priority 2**	
	Layout	1	Layout	1
	Navigation	2		
	Handling	1		
	Wording	1		
	System	1		
Low	**Priority 3**		**Priority 4**	
	Layout	3	Layout	3
	Navigation	1	Navigation	1
	Handling	2	Handling	1
	Data Economy	1	Wording	3

Task 3. Task three is dedicated to the purchase of tickets and their storage on different storage media. Storage on the different media via an app is not yet practice. Therefore, the test subjects expect some problems with these unfamiliar actions. However, as a result, subjects do not perceive the processing to be as challenging as expected. That leads to the positioning of the task in the matrix field "Good Option" (cf. Fig. 2). The expected difficulty level's measurement values scatter significantly more around the mean with 0.98 than the perceived one's values with 0.65. The more homogeneous evaluation of the perceived difficulty level underlines a positive user experience. That offers potential for external customer communication. The new ticket storage options can lead to a positive user experience.

In total, there are 43 interactions between the test moderator and subjects. In 15 cases, the test moderator had to intervene for successful task completion. The tests detect a total of twelve usability problems in the layout, wording, and handling fields of action. The fixing of the problems should be prioritized according to Table 4 and corrected to ensure a good user experience.

Table 5. Task 3 - detected usability problems.

Dimension 2: Influence of problems or errors on the self-reliant task processing	**Dimension 1: Influence of the test subjects' performance level plus the occurred error frequency**			
	High		**Low**	
	Field of action	Problems	Field of action	Problems
High	**Priority 1**		**Priority 2**	
	Layout	1	Wording	2
	Handling	1		
	Wording	1		
Low	**Priority 3**		**Priority 4**	
	Layout	1	Layout	1
	Wording	1	Wording	2
			System	2

Task 4. Task four involves the processing steps of showing and deleting tickets. These are usage scenarios implement with different ticket types and storage media. On average, test subjects expect easier task processing than they perceive at the end, both on a lower value than in test task two. That leads to the positioning of the task in the matrix field "Improve immediately" (cf. Fig. 2). The measurement values of the expected difficulty level of 0.69 scatter around the mean value in a similar range as the values of the perceived difficulty of 0.64.

In total, there are 30 interactions between the test moderator and subjects. In seven cases, the test moderator had to intervene for successful task completion. The tests detect eleven usability problems in the layout, wording, and handling fields of action. The usability of these use cases, especially handling the smartcard and the layout and handling to delete tickets, must be improved before launching the app on the market. The fixing of the problems should be prioritized according to Table 6 and corrected to ensure a good user experience.

Table 6. Task 4 - detected usability problems.

Dimension 2: Influence of problems or errors on the self-reliant task processing	**Dimension 1: Influence of the test subjects' performance level plus the occurred error frequency**			
	High		**Low**	
	Field of action	Problems	Field of action	Problems
High	**Priority 1**		**Priority 2**	
	Handling	2	Layout	2
Low	**Priority 3**		**Priority 4**	
	Navigation	2	Layout	2
			Navigation	1
			Wording	2

Task 5. The expected difficulty level of deleting or removing the user account from the device and uninstalling the app shows is the same value as the perceived one on

average across all subjects. Therefore, task five lies precisely at the intersection of all four strategies (cf. Fig. 2). In total, there are 25 interactions between the test moderator and subjects. In five cases, the test moderator had to intervene for successful task completion.

The expected difficulty level values scatter around the mean of 0.77 and those perceived at 0.89m, both at a relatively high level. These inconsistent ratings concerning the expected and perceived level of difficulty indicate the potential for improvement concerning the ten detected usability problems in fields of action: layout, navigation, wording, handling, and system. The fixing of the problems should be prioritized according to Table 7 and corrected to ensure a good user experience.

Table 7. Task 5 - detected usability problems.

Dimension 2: Influence of problems or errors on the self-reliant task processing	**Dimension 1: Influence of the test subjects' performance level plus the occurred error frequency**			
	High		**Low**	
	Field of action	Problems	Field of action	Problems
High	**Priority 1**		**Priority 2**	
	Navigation	2	Layout	1
	System	1		
Low	**Priority 3**		**Priority 4**	
	Layout	1	Layout	1
	Navigation	2	Wording	1
	Handling	1		

Correlation of the number of detected usability problems and the moderator's interactions. A further step examines the correlation between the test moderator's number of interactions and the usability test's problems. Following the think-aloud method's methodological principles, the moderator should reduce his interactions to a minimum to enable independent task processing.

The analysis of the test results for the TicketIssuance app shows that most interactions (average 5.3 per test subject) between test moderator and test subjects took place in task two. The fewest interactions occurred in Task one, with an average of 0.6 per test subject. The number of interactions in tasks 4 and 5, with an average of 2.7 and 2.3, are similar. In task 3, there is a significantly higher number of interactions with an average of 4.3 per test subject.

The examination of the correlation concerning the five test tasks via the correlation coefficient of Bravis and Pearson results in a correlation coefficient of 0.9438. It shows a strong linear correlation between the average number of interactions per subject and the number of errors detected. The Spearman's rank correlation coefficient with the value of 1.0 shows a robust monotonic correlation, i.e., the more problems or errors occur, the more interactions are required.

Results of Post-Questionnaire. The post-questionnaire responses show that four of the eleven test subjects are very satisfied (rating 1), and seven of them are satisfied (rating 2) with the use of the app.

The test subjects tend to attach greater relevance to the option of saving tickets on the smartphone than on the smart card. Nine out of eleven test subjects rate the option of saving tickets on the smartphone as very important. However, most test subjects also consider the additional option of using an external smart card to be an essential alternative. Four mentions, each of very important and important, clearly show this.

The prerequisite for saving tickets via the app on the SE and an extern smart card is an NCF-enabled smartphone. Seven test subjects state that they have smartphones equipped with NFC; two do not know, and two others answer this question negatively. The test subjects should indicate whether they regularly use mobile payments (e.g., Apple Pay, Samsung Pay), verifying NFC-equipped smartphone knowledge. Five test subjects answered this question in the affirmative, with one of these subjects stating that they do not own an NFC-enabled personal smartphone. It shows that customers often use mobile services as a matter of course without knowing the technical background systems and their devices' technical features.

6 Discussion

6.1 Implemented Measurement and Interpretation Methods

The subjects did not get information about the recording of the time taken to complete the introducing task. These times were the basis for determining each subject's performance level. Clarification about this could have potentially altered the results. The premise was that the test subjects get a practical and straightforward introduction to the test method. The recorded video material analysis from the individual tests formed the basis for determining the number of interactions between test subjects and the moderator and deriving the categories for problems and errors. On the one hand, the researchers' assessment may have led to measurement errors. On the other hand, there is room for interpretation in the problem analysis so that the results are not free of subjective influences.

6.2 Usability Tests Results

In total, eleven test subjects uncovered 56 usability problems and errors in the five test scenarios.

The developed evaluation methodology provides the framework for assigning the collected problems and errors to prioritize the required adjustments categorized in handling, wording, layout, navigation, system, and data economy.

In particular, there is a prior need to adjust the handling and navigation problems. Here, the test moderator often had to interact and intervene to complete the tasks successfully. It was challenging to find central functions such as initializing or configuring the Secure Element or selecting the various storage options. At present, this option is mainly the app's unique selling point and should therefore be intuitive and error-free to use. Potential users are not yet familiar with corresponding operating steps from other applications. It means that they cannot fall back on operating routines. Besides, there is a need for improvement in the layout and wording. The terminology used and the language used for explanations and notes have led to irritation.

6.3 Influence of interactions by the test moderator

The calculated correlation coefficients show a strong correlation between the number of average interactions between the test subject and test moderator and the determined number of usability problems and errors per task. On the one hand, this may be since the test moderator influenced the problem and error detection due to unintentional suggestions. On the other hand, the results may also be due to the moderator's instructions. It required mandatory intervention in specific test scenarios to complete all the tasks successfully.

6.4 Sample Size

The usability test results make it possible to draw concrete conclusions about a suitable number of test subjects who uncover a certain proportion of the existing usability problems and errors. The following formula ($\gamma = 1 - \sqrt[i]{1 - \frac{Foud\ (i)}{N}}$ (3) calculates the proportion of usability problems uncovered by a tester (λ).

$$\gamma = 1 - \sqrt[i]{1 - \frac{Foud\ (i)}{N}} \qquad (3)$$

Thereby, for the present study with 56 uncovered problems by eleven test subjects, an average λ of 0.266 results. Accordingly, each test subject uncovered about 27% of all problems or errors during the usability test. This percentage is about five percentage points lower compared to Nielsen [20]. As Nielsen [20] and Virzi [27] specify, a sample size of five test subjects can elicit about 79% of the problems. The further correlations regarding the influence of the number of test subjects on the percentage of detected errors illustrate Fig. 3. Also shown is the influence at λ = 0.176, calculated based on two randomly selected test subjects, on the proportion of errors detected as a sample size function. It shows that together they uncover just under two-thirds, i.e., 18 out of 56 problems or errors. In the worst-case scenario, five subjects could find (62%), and eleven subjects could find 49 (88%) problems and errors.

The largest λ (= 0.385) calculation is possible with three randomly selected test subjects who uncover 43 of the total 56 items. In this best-case scenario, five subjects uncover 91%, and eleven subjects uncover 97% of the critical points.

For the TicketIssuance app usability study, this means that eleven subjects uncovered 88% to 97% of all problems or errors.

This calculation approach can determine the required number of test subjects for further usability tests to improve the TicketIssuance App. If further testing occurs as part of an iterative design process, it may be advantageous to recruit fewer subjects considering the cost-benefit ratio. According to the worst-case scenario, as few as five ones can uncover more than 60% of all problems and errors. However, if only one more usability test of the app will occur, twelve testers should uncover at least 90% of all problems. Ongoing technological developments usually require continuous adjustments to the app. New problems may arise and require usability testing to ensure user acceptance and a good user experience.

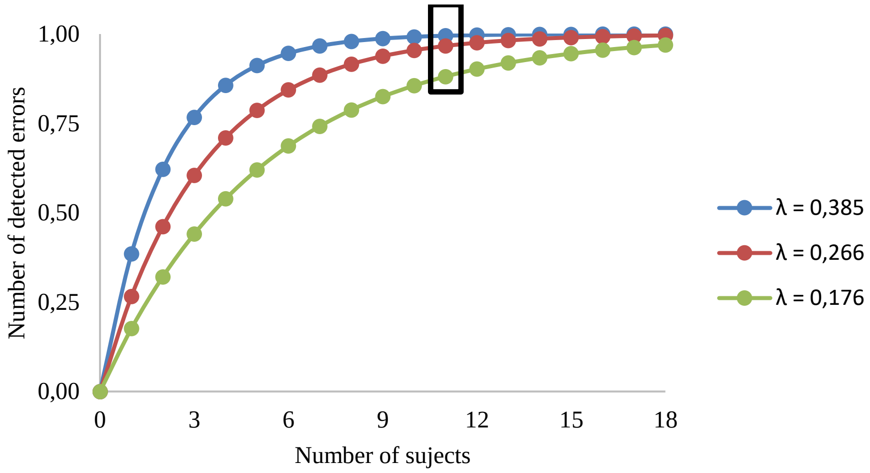

Fig. 3. Results: determination of sample size.

7 Conclusion

The OPTIMOS 2.0 project develops the TicketIssuance App for storing high-price tickets on a high-security level. For this purpose, the app offers new, not yet known functions. The study tested the usability of those new functionalities.

All study subjects completed test tasks. Sometimes the test moderator had to give support. Study subjects perceived the process of ticket purchase easier than expected before testing this function. The satisfaction with the new features for buying tickets and saving them on different external media like a smartcard was very high. Potential users very welcome the possibility of checking smartphone tickets in low battery mode. It was possible to elicit many indications and recommendations for prioritizing of redesigning, particularly regarding the layout, wording, navigation, and handling. The results of the usability tests give suitable bases for further adjustments. There is no need for changing the conception or design of the TicketIssuance app fundamentally, but further tests seem to be necessary. A field test with a larger user group should examine the functional capability of the implemented OPTIMOS 2.0 technology in connection with user-friendliness. Particular attention should lay to the handling of SE configuration and uninstalling processes. The goal must be to ensure a high acceptance level combined with a satisfying user experience for a successful market launch. It can ensure an intensive use of the TicketIssuance app in the future, particularly against the background that the demand for mobile electronic tickets will continue to increase.

References

1. Albert, W., Dixon, E.: Is this what you expected? The use of expectation measures in usability testing. In: Proceedings of the Usability Professionals Association 2003 Conference, Scottsdale, AZ (2003).
2. Albert, W.; Tullis, T. Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics. Morgan Kaufmann, Burlington, MA, USA, 2013
3. Alhadreti, O., Mayhew, P.: Rethinking Thinking Aloud: A Comparison of Three Think-Aloud Protocols. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 1–12 (2018).
4. Barnum, C. M.: Usability Te sting Essentials: Ready, Set. Test! Elsevier, Inc., Burlington, MA, (2011).
5. Berliner Verkehrsbetriebe: BVG Lagebericht & Jahresabschluss 2018. https://unternehmen.bvg.de/wp-content/uploads/2020/10/BVG-Lagebericht-2018.pdf, last accessed 2021/02/23.
6. Bevan, N.: Classifying and selecting UX and usability measures. In: International Workshop on Meaningful Measures: Valid Useful User Experience Measurement, Volume 11, pp. 13–18, (2008).
7. Bilandzic, H., Trapp, B.(2000): Die Methode des lauten Denkens: Grundlagen des Verfahrens und die Anwendung bei der Untersuchung selektiver Fernsehnutzung bei Jugendlichen. In: Paus-Haase, I., Schorb, B. (eds.): Qualitative Kinder-und Jugendmedienforschung, pp.183–209, KoPäd - Kommunik. u. Päd., München (2000).
8. Ericsson, K.A., Simon, H.A.: Protocol analysis: Verbal reports as data. 2nd edn. MIT Press, Cambridge, MA, (1993)
9. Fonteyn, M. A., Kuipers, B., Grobe, S.J. (1993): A Description of Think Aloud Method and Protocol Analysis. In: Qualitative Health Research 3 (4), pp. 430–441, (1993).
10. Frühwald, A. (1955): Das REFA-Gedankengut – Eine Darstellung für den Kaufmann. Gabler, Wiesbaden (1955).
11. Gerike, R., Hubrich, S., Ließke, F., Wittig, S., Wittwer, R.: Tabellenbericht zum Forschungsprojekt „Mobilität in Städten – SrV 2018“ in Berlin. Technische Universität Dresden (2020).
12. Harrison, R., Flood, D., Duce, D.: Usability of mobile applications: literature review and rationale for a new usability model. In: Journal of Interaction Sciences, 1 (1), pp. 1–16 (2013).
13. Henry, S. B., Lebreck, D. B., Holzemer, W. L.: The effect of verbalization of cognitive processes on clinical decision making. In: Research in Nursing & Health, 12 (3) pp. 187–193 (1989).
14. Jakobitz, D., Krüger, S. (2018): Tickets auf dem Smartphone - das wünschen sich Fahrgäste im ÖPNV. Unter: https://marktforschungsanbieter.de/files/profiles/249/12387354437530.pdf, last accessed 2021/02/23.
15. Kaiser, M. O: Erfolgsfaktor Kundenzufriedenheit – Dimensionen und Messmöglichkeiten. 2nd edn. Erich Schmidt Verlag, Berlin: (2005).
16. Lewis, J. R.: Sample sizes for usability studies: Additional considerations. Human factors 36 (2), pp. 368–378 (1994).
17. Nielsen, J.: Heuristic Evaluation. In Nielsen, J., Mack, R. L.: Usability Inspection Methods. pp 25–62, John Wiley & Sons, New York (1994).
18. Nielsen, J.: Usability Engineering. Academic Press Inc., Cambridge (1993)

19. Nielsen, J., Landauer, T. K.: A mathematical model of the finding of usability problems. In: Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems, pp. 206–213 (1993).
20. Nielsen Norman Group: Why you only need to test with 5 users. https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users, last accessed 2021/02/23.
21. Nielsen Norman Group: Thinking Aloud: The #1 Usability Tool. https://www.nngroup.com/articles/thinking-aloud-the-1-usability-tool, last accessed 2021/02/23. Olmsted-Hawala, E. L., Murphy, E. D, Hawala, S., Ashenfelter, K. T.: Think-aloud protocols: a comparison of three think-aloud protocols for use in testing data-dissemination web sites for usability. In: Proceedings of the SIGCHI conference on human factors in computing systems, pp. 2381–2390 (2010).
22. Olmsted-Hawala, E. L., Murphy, E. D, Hawala, S., Ashenfelter, K. T.: Think-aloud protocols: a comparison of three think-aloud protocols for use in testing data-dissemination web sites for usability. In: Proceedings of the SIGCHI conference on human factors in computing systems, pp. 2381–2390 (2010).
23. Rhenius, D., Deffner, G.: Evaluation of concurrent thinking aloud using eye-tracking data. In: Proceedings of the human factors society annual meeting, Volume 34 (17), pp. 1265–1269 (1990).
24. Schlittgen, R.: Einführung in die Statistik – Analyse und Modellierung von Daten. 8th edn., Oldenbourg München Wien (1998).
25. Spool, J., Schroeder, W.: Testing web sites: Five users is nowhere near enough. InCHI'01 extended abstracts on Human factors in computing systems, pp. 285–186 (2001).
26. Tergan, S. O.: Grundlagen der Evaluation: ein Überblick. In: Schenkel, P., Tergan, S. O., Lottmann, A. (eds.): Qualitätsbeurteilung multimedialer Lern- und Informationssysteme – Evaluationsmethoden auf dem Prüfstand. Pp. 22–51. BW Bildung und Wissen, Nürnberg (2000).
27. Virzi, R. A.: Streamlining the design process: Running fewer subjects. In: Proceedings of the Human Factors Society Annual Meeting. Volume 34 (4), pp. 291–294 (1990).
28. YouGov: App in die Tonne – Wie Ihre App geladen, getestet und schließlich nicht wieder gelöscht wird. YouGov Reports, Köln (2017).