Journal of Ubiquitous Systems & Pervasive Networks
Volume 9, No. 1 (2017) pp. 01-10

Safety-Critical Mobile Systems – The RESCUER Interaction Evaluation Approach

Konstantin Holl[a*], Claudia Nass[a], Vaninha Vieira[b], Karina Villela[a]

[a]*Fraunhofer IESE, Kaiserslautern, Germany, 67663*
[b]*Fraunhofer Project Center, Federal University of Bahia, Salvador, Brazil, 40110-170*

Abstract

The infrastructure organization of large-scale events involves high safety requirements for the visitors and is a central issue for the officials in charge. To assist in dealing with this, we developed the RESCUER Mobile Crowdsourcing App, which runs on smartphones and allows the crowd to report an emergency, thereby improving the process for rescuing humans in an emergency. For the evaluation of the app, we faced the problem that people participating in a large event, such as a soccer match, are not willing to spend time on completing a long survey or interview. Also, people experiencing an emergency situation may have their cognitive capabilities affected by emotional burden, so a mobile app should be easy and intuitive to interact with. Hence, the goal of this contribution was to select and perform an on-site mobile evaluation approach that allows us to evaluate the user interaction. Two main evaluations were performed using two different versions of our application. The first evaluation took place during the FIFA World Cup 2014 and tested the app's usability with 112 users in Brazil and in Germany. As a result of this evaluation, we found severe usability issues and gained concrete insights into how to solve them. The second, follow-up evaluation, using an improved version of our app, was performed during emergency exercises in Brazil, with 31 experts in emergency management. For our evaluation approach, the results indicated that on-site mobile evaluation is an appropriate method for improving the usability and interaction of safety-critical software systems.

Keywords: *Mobile app; Emergency management; Evaluation; Usability*

1. Introduction

The measurement of usability and user experience as part of user studies is made possible by the use of quantitative and qualitative methods [1]. These methods can be applied as part of controlled environments as well as in real contexts. However, there are important questions regarding the adequacy of these methods. Lang (2013), for example, claims that the usage of user studies in the field of mobile software systems within a real context of use is overrated and is not necessary [2]. This statement is based on statistical insights such as that 73% of smartphones users also use their devices in bed, respectively on the couch. This implies that user tests in controlled environments are more adequate because there would be reproducible influence factors. Contradicting Lang (2013), IGD (2015) revealed that 69% of people use their smartphones during work, 51% on weekends, and 42% while commuting between home and work [3]. This means that a preference for controlled environments for evaluation purposes would imply that many usage contexts would not be taken into account.

For user evaluations test, it is essential to understand a user's behavior in a real context, such as while interacting with objects of daily use. This includes i) considering the fulfilment of user requirements in specific usage scenarios and ii) analyzing the extent to which the system meets the mental model of use [1].

These objectives can only be achieved by field studies. In contradiction to Lang (2013), we consider it highly relevant to evaluate mobile software systems in real settings as part of field studies [2]. Known approaches for field studies of mobile software systems are, for instance, Guerrilla and Lightweight Testing [4], In-The-Wild Testing [5], or Fly on the Wall-Study [6]. These types of studies show that evaluations in the field – i.e., where the users apply the mobile software system – reveal deep insights into the usage of the mobile software system. The performance of user tests in real contexts may be quick and flexible. Furthermore, they are adequate for samples of any size and enable the combination of quantitative and qualitative measurement approaches in order to examine the interaction between mobile devices and the real world [7]. These factors were essential arguments for our decision to perform field studies in order to collect experiences.

The field study presented in this paper was performed within the scope of the project RESCUER (Reliable and Smart Crowdsourcing Solution for Emergency and Crisis

* Corresponding author. Tel.: +49 631 6800 2284
Fax: +9876543210; E-mail: konstantin.holl@iese.fraunhofer.de

DOI: 10.5383/JUSPN.09.01.001

Management) [8], which aims at developing a smart platform to support communication between people at the place of an incident and people at a command center, based on reliable and intelligent analysis of crowdsourcing multimedia information in two scenarios: during large-scale events and in industrial parks. The RESCUER Mobile Crowdsourcing Solution (MCS) is one of the components of the RESCUER platform, illustrated in Figure 1, which also includes the Data Analysis Solution, the Emergency Response Toolkit (ERTK), and the Ad-hoc Communication Solution. RESCUER MCS (or MCS for short, as used in the following) supports the notification and characterization of an emergency situation by involving the crowd at the place of an incident. Our goal was to explore the participation of the crowd just after an incident has occurred, while they are still close to the site of the incident, in terms of informing the command center about the incident and describing its main characteristics. Consequently, we needed to answer the following question:

> How to evaluate the usability and interaction of a mobile application for emergency situations at the site of a large-scale event or an industrial park?

Particularly right before a large event, such as a soccer match, visitors are not willing to spend time on completing a long survey or interview. Similarly, volunteers and experts participating in an emergency exercise are involved in several tasks and are not willing to spend time. Therefore, when planning the performance of a user field study, it is necessary to keep it as short as possible while still covering all essential aspects. This paper describes the RESCUER MCS, our on-site evaluation approach, and its results, and discusses these. Finally, we present insights and lessons learned based on two studies: the first performed at venues of the FIFA World Cup 2014, and the second during emergency exercises in Brazil.

RESCUER MCS supports the communication of eyewitnesses and official first responders (e.g., police, firefighters) with a command and control center. It is therefore also referred to as the RESCUER app or RESCUER application. Eyewitnesses and first responders are equipped with variations of MCS for the following purposes:

- Eyewitnesses use MCS on their mobile devices to provide and receive information about an incident that has occurred. The goal is to benefit as much as possible from information that can be provided by mobile devices without any explicit action of their users, but taking into consideration the user's privacy; and
- First responders should first focus on rescuing victims, providing medical care, and dealing with hazards. They will mainly use mobile devices, such as smartphones and wearable devices, equipped with MCS to keep the command center informed about the evolution of the situation, ask for reinforcements, and receive instructions.

Based on that, two types of information can be gathered from people carrying mobile devices at the place of an emergency situation: (1) information that can be extracted from mobile devices without user interaction with those devices (e.g., GPS position, movement speed, movement trails, and number of devices at a specific location), as well as (2) information to be provided by users through explicit interaction with their mobile device (e.g., videos and images of the incident, text messages describing the incident).

One must consider that emergency situations, where people are often confused and overwhelmed, trigger very basic instinctive human behavior. Therefore, it is important to understand how these situations affect the human behavior. Having a mobile device in an emergency situation offers the possibility for promptly asking for help, but, depending on the specific circumstances, the user might be distracted from using the mobile device or may accidentally use it in the wrong way.

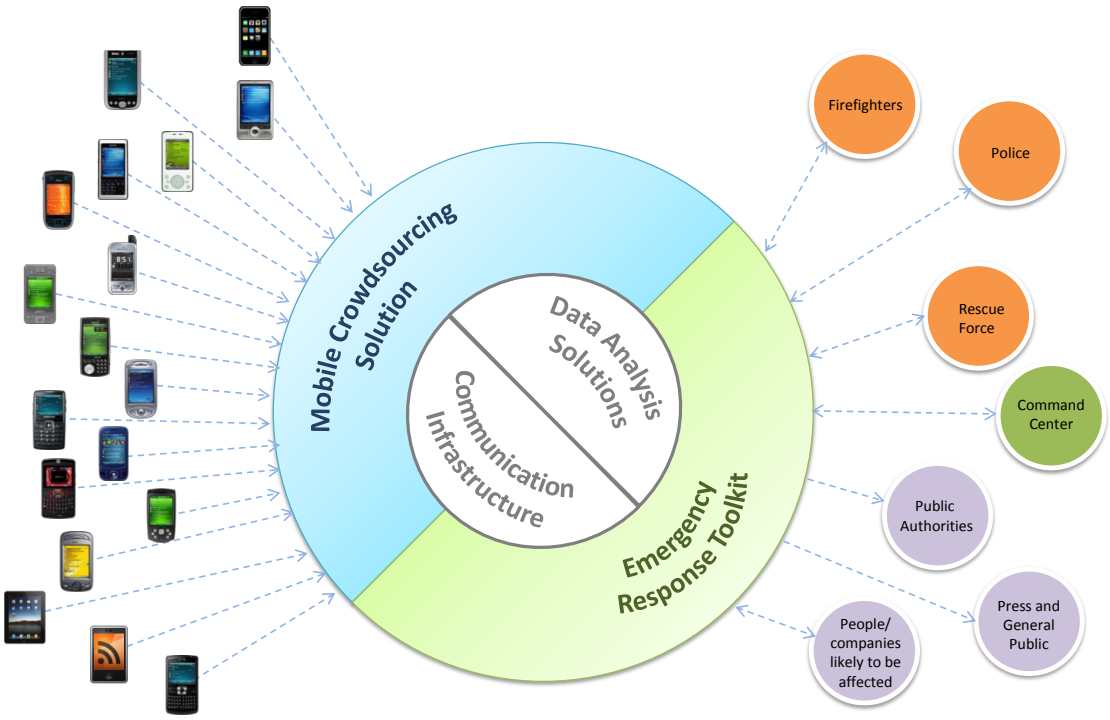


Fig. 1. Interaction model of the RESCUER application user interface

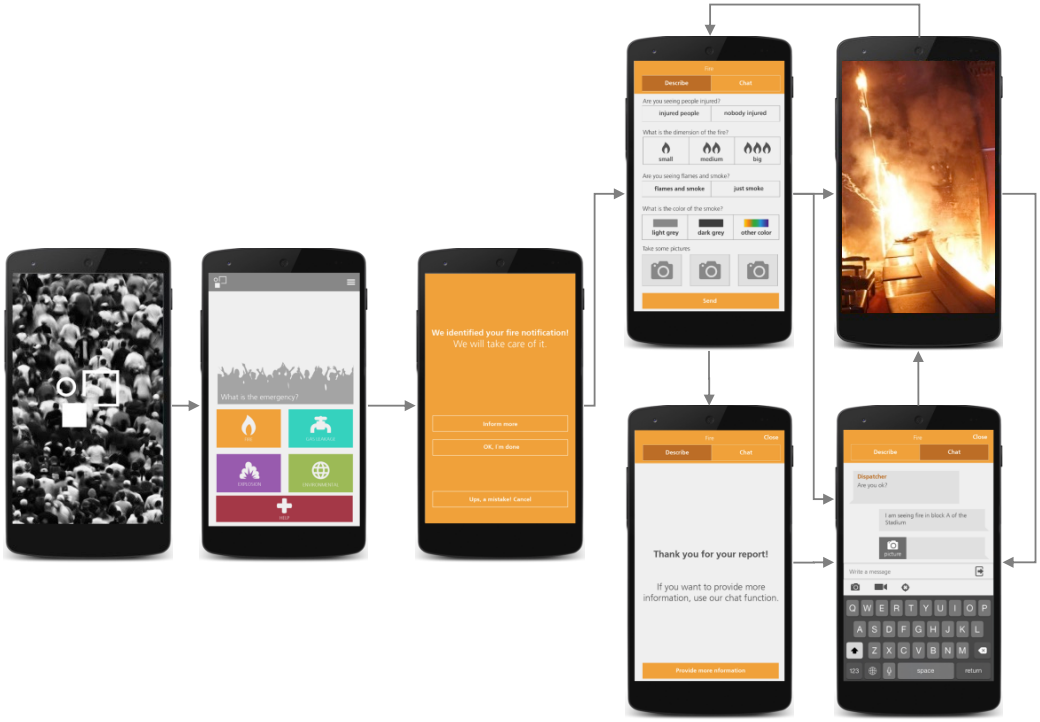


Fig. 2. Interaction model of the RESCUER application user interface

2. The RESCUER Mobile Crowdsourcing Solution

Figure 2 presents the interaction model of MCS's user interface, which is based on three types of interaction allowing eyewitnesses to provide information according to their needs: one click, guided, and free text. In short, users start the report of an incident by notifying the command and control center about the type of the incident, such as an explosion, a fire, or human crush. By pressing one of the options, the reporting process is immediately triggered and sensor-based information is sent to the server (one-click interaction). After this, if they move to a safer place or feel less emotionally affected, users can continue the interaction and send a standard report or recall the notification in the case of a false alarm. In a standard report, it is possible to specify the severity of the incident, state if there are injured persons, and take photos/videos of the incident, if desired (guided interaction). In addition, the users can send free text messages to provide more specific information or send the exact position of the incident on a map once they feel safe enough to do so (free text interaction).

By filtering, combining, and analyzing different pieces of crowdsourcing information, the emergency command center should be able to react to an emergency more quickly and efficiently. To help achieve this goal, the RESCUER app should have a high degree of usability and support the user even in stressful situations with a user-friendly design. The appendix shows a possible scenario considering the interaction between the RESCUER app and the ERTK (see Figure 12).

Figure 3 gives an overview of the roles that are considered by the RESCUER system. The Appendix shows an illustrated process that comprises the RESCUER app and the ERTK component considering these roles. Roles include MCS users, who are represented by eyewitnesses, first responders, and operational forces. The command and control center is the main role of the ERTK. The affected community and the general public are users of the RESCUER solution, which is integrated into the ERTK.

Fig. 3. RESCUER roles

3. The RESCUER On-Site Evaluation Methodology

The on-site evaluation performed in RESCUER took place in the context of the FIFA World Cup 2014 in Brazil and in Germany in three different cities, namely Salvador and São Carlos (Brazil) and Kaiserslautern (Germany). Because of the nature of the RESCUER app (support during emergencies at large-scale events), it was relevant for the on-site evaluation to choose a location where a large-scale event was taking place. The evaluation sites were the public viewing areas in these cities. There, we were allowed to perform our study in the two hours before the matches started. In total, 112 people participated in the on-site evaluation.

Two follow-up evaluations were performed in June 2016 with the new version of the RESCUER app, following improvements made as a result of feedback received in the first evaluation: a) during II CIDEM (International Mass Disaster Conference) in Salvador (Brazil); and b) during a fire exercise simulation at the Industrial Park of Camaçari, in Camaçari (Brazil), with employees of the Park and COFIC. These evaluations will be discussed in Section 6.

Fig. 4. Stimuli used for the description of the incident (photo extracted from Turner (2009) [10]

The main purpose of the on-site evaluation was to assess the usability of the RESCUER app in the context of use. One important aspect to be considered in such a setting is the short duration of the evaluation as well as the mobility of the participants and the evaluation team. For this reason, a mix of measurement methods was prepared for the evaluation. Table 1 shows the detailed measurement goals and methods used.

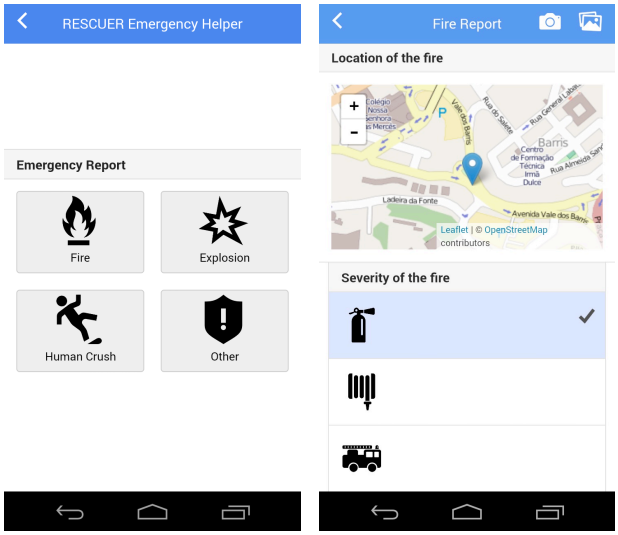


Fig. 5. Emergency selection via RESCUER app (left) and Fire Report Screen (right)

A person acting as coordinator performed the first approach to identify potential participants and randomly select the people who were to perform the evaluation. The users were divided into two groups: (i) users who performed the tasks without previous demonstration of the RESCUER app; and (ii) users who performed the tasks after a short demonstration. This division into groups was very important, as it allowed us to check the learnability of the RESCUER app.

The RESCUER app evaluation was supported by two people: a moderator and an observer. The moderator was responsible for addressing the participants, presenting the application, supporting the participants during the test, and applying the survey. The observer was responsible for filling out the observation sheet and handling the evaluation cards.

The moderator showed a picture of a fire in a stadium (see Figure 4) and asked the participants to imagine that the fire observed in the picture was burning on the other side of the stadium or venue and that they wanted to inform the firefighters by means of the RESCUER app (see Figure 5). After this instruction, the moderator assigned the tasks to be performed by the participants. For each task, the observer completed the following items: Did the user accomplish the task successfully? Did the user accomplish the task without further questions, after 1-2 questions, after 3-4 questions, or after 5 or more questions?

After performing all assigned tasks in the application, the users were asked to complete a Mini AttrakDiff questionnaire and a demographic data survey. AttrakDiff is an established evaluation tool that addresses evaluations of user experience and has already been used for evaluating mobile systems [9]. It consists of pairs of contrasting attributes that can be applied to the application. The squares between the attributes represent gradations between the opposites. The user can express his/her agreement with the attributes by checking the box that most closely reflects his/her impression.

All information (observer notes and participant's answers to the questionnaire) related to one participant was recorded on one single evaluation card, as presented in Figure 6.

Table 1. Measurement goals and methods

Measurement Goal	Measurement Method
Effectivity	Number of users who accomplished the given tasks: (1) Report that you see a fire (2) Provide information that the area of the fire is on the other side of the stadium (3) Provide information on whether you see injured people (4) Describe the properties of the fire (5) Provide information about the severity of the fire (6) Take a photo of the fire
Pragmatic and hedonic quality	Mini AttrakDiff
Demographic data	Personalized questionnaire (gender, age, own smartphone).
Improvement potential	Notes of the evaluation team

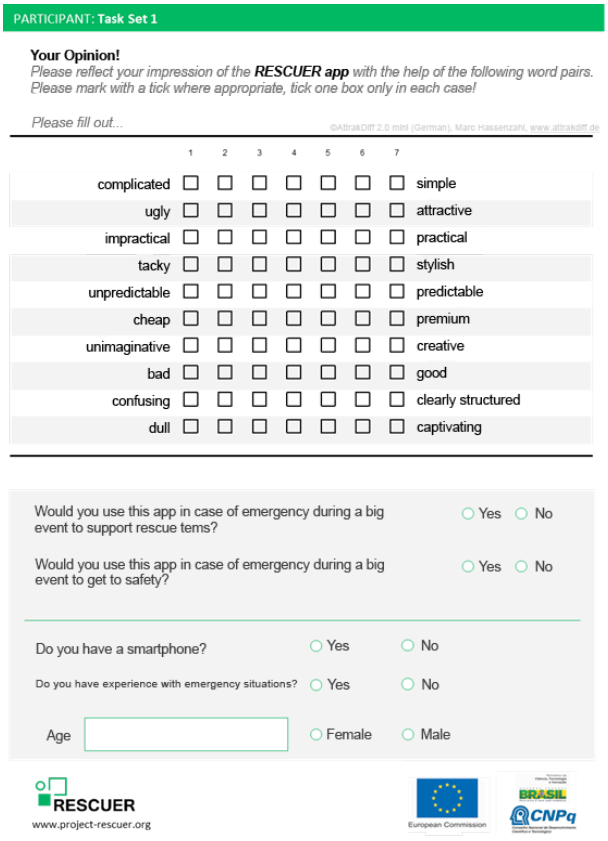


Fig. 6. Evaluation card used for the study participants

4. Results of Study 1: FIFA World Cup 2014

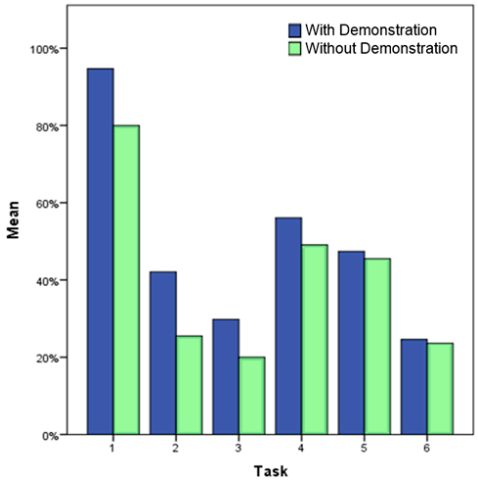


Fig. 7. Participants who accomplished all tasks with a demonstration

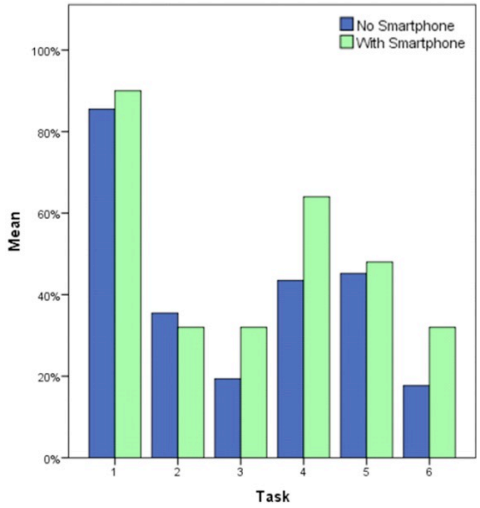


Fig. 8. Participants who accomplished all tasks and owned a smartphone

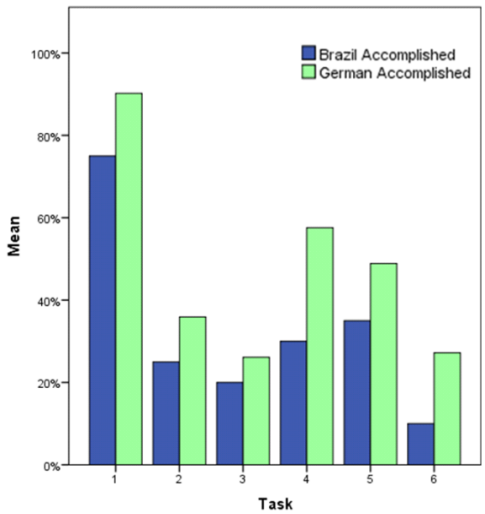


Fig. 9. Participants who accomplished all tasks considering the location

The participants of the study consisted of 64 male and 48 female visitors of the public viewing areas of the FIFA World

Cup 2014 (N = 112). The participants' age ranged from 13 to 68 years (M = 26.21, SD = 11.67). 72.4% of the participants in the evaluation owned a smartphone. In summary, 50 participants were involved in the evaluation in Kaiserslautern (Germany), 35 in Salvador (Brazil), and 27 in São Carlos (Brazil). Of these, 49.1% used the RESCUER app without any demonstration and 50.9% with a short demonstration.

First, we present the results related to the effectiveness of the RESCUER app, i.e., the number of users who performed a task successfully. A successfully performed task means that the participant accomplished the task without further questions about the interaction or use of the RESCUER app. Then we present the results of the users' subjective assessments related to the app's pragmatic and hedonic qualities by means of the Mini AttrakDiff questionnaire.

Task 1 (report that you see a fire) was successfully performed by 83% of all participants, whereas the subsequent two tasks (2 – Provide information that the area of the fire is on the other side of the stadium; 3 – Provide information on whether you see injured people), and the last task (6 – take a photo of the fire) were accomplished by only about one out of four participants. Tasks 4 (Describe the properties of the fire) and 5 (Provide information about the severity of the fire) were accomplished by every second participant.

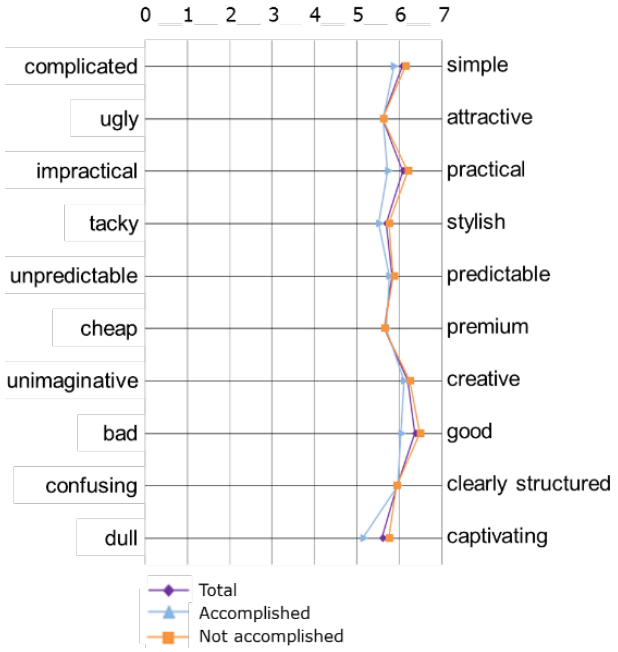


Fig. 10. Mini AttrakDiff survey results

The relative number of participants who accomplished all tasks and had received a demonstration before is slightly higher than that of the participants who accomplished all tasks and had not received a demonstration before. Especially for the first three tasks, this was significant, due to the particular differences ranging between 9% and 15% (see Figure 7). Overall, the results show that prior demonstration supported the accomplishment of the tasks and increased the likelihood of accomplishing all tasks successfully. The significance test was done based on the phi coefficient, respectively the mean square contingency coefficient ($\phi = -.254$, $p < .01$).

Not every participant owned a smartphone. Consequently, we compared the relative number of participants who accomplished all tasks and owned a smartphone to that of the participants who accomplished all tasks and did not own a smartphone. Considering the maximum difference regarding Task 4, 10% more participants owning a smartphone accomplished the task compared to participants without a smartphone. An exception were the results regarding Task 2, which indicate that there were slightly more participants who could accomplish this task and did not own a smartphone. Nevertheless, Figure 8 implies a slight tendency that smartphone possession supports the accomplishment of the tasks. Considering our significance test, the successful accomplishment of all tasks did not depend on whether a participant owned a smartphone or not ($\phi = -.116$, $p=.22$).

The evaluation was performed in Brazil and Germany. The results presented in Figure 9 show the differences in success regarding the accomplishment of all tasks by participants in Brazil and participants in Germany. The differences range from 5% to 27% (see Task 3 and Task 4 in Figure 9). Overall, the rate of task accomplishment considering all six tasks was higher in Germany than in Brazil. The percentage of German participants who accomplished all tasks successfully seems to be slightly higher than that of the Brazilian participants. However, the difference does not reach significance ($\phi = -.116$, $p = .08$).

Regarding the measurement of the pragmatic and hedonic qualities of the user interface, the feedback of the participants was generally positive. The average score was 6 out of a maximum of 7 for almost all attribute pairs. Figure 10 shows three views of the Mini AttrakDiff results.

The total view shows the average graph of all participants, while the accomplished view shows only the average graph of the participants who performed every task successfully. The not accomplished view shows the average graph of all the participants who did not fulfill at least one of the tasks or who required some support from the experimenters. Overall, the differences among the three views are very slight. The average graphs are invariably located between the score points 5 and 7. The maximum score distance between all graphs is 0.7 points.

4. Follow-Up Evaluation – A Quantitative Analysis

The follow-up evaluation was conducted during II CIDEM (in English: International Conference on Mass Disasters). This event was attended by members of local, regional, national, and international organizations that act in disaster situations, such as Civil Defense, Municipal Guard, Firefighters, Military Police including the Special Operations Police (BOPE), Civil Police, Technical Police, Federal Police, Army, Navy, Air Force, Interpol, and universities. This event took place in Salvador, Brazil, at the Hotel Fiesta and the Arena Fonte Nova, from 10 to 12 June 2016; the theme was "Safety for large events – A global warning".

In this evaluation we used the lobby of the hotel, an area next to the conference room with seating for 1,000 people. In the evaluation we tried to transfer the participants to the scenario of a fire in the conference room. We used posters with images and texts to make the evaluation participants of each profile (civilian, supporting force, or workforce) feel like they were on their way from the hot zone to the cold zone.

In total, 31 people participated in this evaluation, distributed across three profiles: civilian, supporting force, and workforce. In addition, supporting forces and workforces informed their sub-profiles. Regarding their experience in emergency situations, most of the supporting forces and workforces had some experience in emergency situations. However, 54% of the civilians said they had not had any experience in emergency situations. Regarding the gender of the participants, 57% of them were male, whereas 43% were female.

The quantitative analysis of the evaluation results takes into consideration the quality characteristics and sub-characteristics mentioned above. Table 2 summarizes the relationships between questions used as data sources quality characteristics and sub-characteristics, metrics, and acceptance criteria.

For each quality characteristic/sub-characteristic, the quantitative results were calculated based on the metrics specified in Table 2. In the case of more than one question contributing to a quality attribute, the final measure is the average of the results of each question.

Table 3 presents the acceptance criterion, final measure, and acceptance result for each quality characteristic/sub-characteristic. As shown in Table 3, two quality characteristics/sub-characteristics, namely Completeness and Freedom from Risk, did not satisfy their respective acceptance criteria.

Table 2. Basis for the quantitative analysis

Question/Assessment	Quality Attributes	Metric	Acceptance Criteria
Q1. Is there any information that you think is relevant for handling a fire incident, but you could not provide it using the app?	Functional Suitability (Completeness)	Number of participants who answered No	80%
Q2. Is there any information that you provided or could have provided, but you would like to have better support from the app?	Functional Suitability (Appropriateness)		
Q3. Is there any part of the app (e.g., taking picture, free text report) that you think is not relevant for reporting an incident?	Functional Suitability (Appropriateness)		
Q4. Is there any piece of information (e.g., color of the smoke) in the report form that you know is not relevant for handling the incident?	Functional Suitability (Appropriateness)		
Q5. Would you use this app to help operational forces if an emergency situation like this occurs at a large-scale event?	Usability (Appropriateness/Recognizability), Usefulness and Trust	Number of participants who answered Yes	80%
Q6. Would you use this app to request help (for yourself) if an emergency situation like this occurs at a large-scale event?	Usability (Appropriateness/Recognizability), Usefulness and Trust		
Q7. I will feel safer with the RESCUER app on my smartphone during a large-scale event.	Usefulness, Trust, Freedom from Risk	Number of participants who answered “Agree” or “Strongly Agree”.	80%
Q8. Did you have the feeling of running further risks when using the app?	Freedom from Risk	Number of participants who answered “In none of the scenarios”.	80%
Q9 to Q17: AttrakDiff Questionnaire	Appropriateness/Recognizability, Operability, Learnability, User Interface Aesthetics	Score obtained in each AttrakDiff criterion	Average score should be > 75%

Table 3. Basis for the qualtitative analysis

Quality Characteristic	Quality Sub-Characteristic	Acceptance Criterion	Final Measure	Acceptance Result
Functional Suitability	Completeness	0.8	0.65	***No***
	Appropriateness	0.8	0.8	***Yes***
Usability	User Interface Aesthetics	0.75	0.84	***Yes***
	Appropriateness/Recognizability	0.75	0.93	***Yes***
	Learnability	0.75	0.83	***Yes***
	Operability	0.75	0.86	***Yes***
Satisfaction	Usefulness	0.8	0,97	***Yes***
	Trust	0.8	0.97	***Yes***
Freedom from Risk		0.8	0.7	***No***

Regarding Completeness, several participants mentioned that they missed the possibility of providing certain information, but almost all missing information can already be provided (e.g., smoke color and number of injured people) or should not be provided by the specific user profile (e.g., incident status by civilians). We also believe that the lack of the follow-up and guidance features at that point in time also influenced the results, as one person asked for general emergency instructions. Concerning Freedom from Risk, as previously discussed, several participants mentioned they had the feeling they were

running risks when providing information in the incident scenarios. We plan to analyze the records of the usage of the RESCUER app by each participant in order to find out if they used the app according to the level of risk posed by the scenario or if they used it excessively because they were not in a real emergency situation. Nevertheless, all considerations presented in the previous sections will be carefully considered in the new versions of the RESCUER application.

Last but not least, Functional Suitability and Usability (product quality characteristics) contribute to the perception of Usefulness, Trust, and Freedom from Risk (quality-in-use characteristics). The fact that the results for product quality characteristics and quality-in-use characteristics, which were calculated independently, are aligned demonstrates the consistency of the evaluation results.

5. Discussion

The results related to effectiveness show severe problems regarding the usability of the RESCUER app, especially with respect to the performance of Task 2 (Provide information that the area of the fire is on the other side of the stadium), Task 3 (Provide information on whether you see injured people), and Task 6 (Take a photo of the fire).

It was possible to observe, and confirmed by the notes taken by the evaluation team, that the users were not able to read the map when performing Task 2. Although the current position was displayed with a pin on the map, people were not able to get their orientation on the map and identify the other side of the stadium. Considering that in a real emergency situation, the level of stress is higher and the interaction with the device has to compete with more important tasks, such as exiting the venue, this does not appear to be an appropriate interaction concept for a mobile app like RESCUER.

When performing Task 3, people had difficulties identifying whether injured people were present in the supporting picture. People could only provide information on whether they saw injured people or not (i.e., they had to decide between two options). If they were unsure whether there were injured people in the picture, they started to ask for an evasive answer. This is a situation that can easily occur in a real emergency. One possible solution could be to ask the participants about the number of injured people they can actually see or to include the option of an evasive answer such as “I don’t know” or “I cannot assess”.

Task 6 could often not be performed successfully by the participants because they were unable to find the camera button at the top of the RESCUER app. Most of them looked for the camera button at the bottom of the graphical user interface and were therefore unable to perform the task successfully.

In the analysis of the effectiveness results, several insights revealed possibilities for improving the RESCUER app. The combination of objective quantitative methods with the comments of the evaluation team, in particular, helped to clarify several usability issues of the app.

Nevertheless, the results of the Mini AttrakDiff show a very positive assessment of the RESCUER app. These results were good in principle, but they were not expected at all when observing the number of people who were unable to operate the RESCUER app seamlessly. Different factors could have contributed to this result: (1) The participants were in touch with the RESCUER app for a really short time (about 2 minutes); during this time, it is difficult to form an opinion about the system; (2) the evaluation situation may have influenced the completion of the questionnaire since the evaluation team was close to the participants, more or less observing them during this time; and (3) people were in a hurry, they wanted to get good places, organize some food or drinks, and did not focus on filling out the questionnaire.

The quantitative analysis complemented these insights. The RESCUER app could be improved especially with regard to the input options for information that the users think is relevant for handling a fire incident. This could improve the completeness of the app. Furthermore, the users are partially afraid of running further risks when using the app. This should be addressed with usability improvements in order to mitigate the low level of freedom from risks felt by the participants.

6. Lessons Learned

In this paper, we presented an on-site evaluation for a crowdsourcing-based mobile app for emergency situations, performed in the context of the FIFA World Cup 2014 with 112 participants and as part of emergency exercises with 31 experts in emergency management in 2016. The evaluations reveal severe usability issues in the RESCUER app and provided insights into on how to solve these problems in order to improve the software system in early stages of its development. This study shows that on-site evaluation can be an appropriate method for evaluating the usability and user experience of a safety-critical system such as RESCUER.

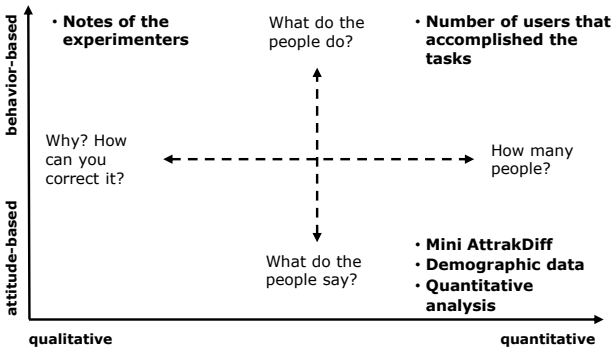


Fig. 11. Classification of the measurement approaches of the RESCUER app field study based on Rohrer (2014)

The main lesson learned for future on-site mobile evaluations is related to the measurement methods applied in such a context. Rohrer (2014) presents a 3-dimensional framework for the systematic selection of UX research methods [11]. Within this framework, several UX research methods are classified into three dimensions:

1. Behavioral (what people do) vs. Attitudinal (what people say)
2. Qualitative (why or how to fix) vs. Quantitative (how many and how much)
3. Usage context (natural or near-natural use of the product, scripted use of the product, not using the product during the study, a hybrid of the others)

In order to verify the quality attributes of the RESCUER app, we selected a mixture of several measurement approaches (see Figure 11) referring to a diversity of data and elicitation methods.

Based on the classification of Rohrer (2014) [11], we conclude that behavioral methods brought more insights related to problems of the RESCUER app as well as ideas for solving these problems than attitudinal methods.

Particularly in the context of a quick on-site evaluation, where the participants interact with the mobile app for a short time, questionnaires do not seem to be appropriate for generating new knowledge regarding improvement of the system.
Furthermore, we learned that a coordinator was necessary to preselect the participants. This was not defined in the original version of the method. Besides, we determined that two people – an observer and a moderator – were a good approach to make the test faster and to make it easier to get insights and take notes.
In general, evaluations regarding mobile apps should lead to the use of simpler and more effective methods that help researchers get a better understanding of how people behave in their normal environment and when interacting with the mobile app.

Acknowledgments

This contribution is funded by the RESCUER project (European grant no. 614154; MCTI/CNPq grant no. 490084/2013-3). The project consists of a partnership between the European Union and Brazil. Leading researchers from renowned institutions and industries joined forces to form this partnership.

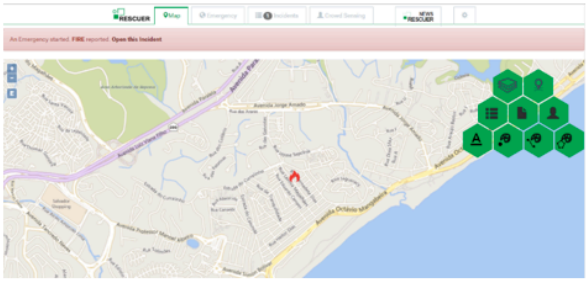
References

[1] Kuniavsky, M. 2003. Observing the User Experience: A Practitioner's Guide to User Research. San Francisco: Morgan Kaufmann Publishers, Inc.
[2] Lang, T. 2013. Eight Lessons in Mobile Usability Testing. Retrieved May 25, 2015 from http://uxmag.com/ articles/eight-lessons-in - mobile-usability-testing.
[3] IGD. 2015. IDG Global Mobile 2014 Survey. Retrieved May 25, 2015 From:http://idgknowledgehub.com/ mobileidg/idg-mobile-survey/.
[4] Mankelow, T. 2012. 10 Tips for Mobile Usability Testing. Retrieved May, 25, 2015 from http://optimalexperience.com/2012/01/10-tips-for-mobile-usability-testing/
[5] Kanneganti, S. 2015. 5 Reasons for Going Remote with Mobile Website Usability Testing. Retrieved May, 28, 2015 from http://www.userzoom.com/mobile-usab ility-testing-blog/5-reasons-for-going-remote-with-mobile-website-usability-testing/#_ftn3.
[6] Optimal Experience. 2015. Field studies. Retrieved May, 25, 2015 from http://optimalexperience.com/what-we-do/field-studies/.
[7] User Zoom. 2015. Mobile In-The-Wild Testing. Retrieved May, 27, 2015 from http://www.user zoom.com/mobile-in-the-wild-testing/.
[8] Villela, K., Breiner, K., Nass, C., Mendonca, M., Vieira, V. 2014. A Smart and Reliable Crowdsourcing Solution for Emergency and Crisis Management. 22nd Interdisciplinary Information Management Talks (IDIMT 2014): Networking Societies -Cooperation and Conflict. Podebrady, Czech Republic.
[9] Bertram, D., Blum, R., Bürger, S., Ecker, M., Henkel, T. and Katsaros, V. 2010. Usability-Testbericht. Hochschule der Medien, Stuttgart.
[10] Turner, L. 2009. A flare for soccer: open flames in soccer stadiums worldwide. Retrieved July, 13, 2015 from http://www.boston.com/sports/blogs/bigshots/2009/ 05/a_flare_for_soccer_open_flames.html.
[11] Rohrer, C. 2014. When to Use Which User-Experience Research Methods. Retrieved May, 29, 2015 from http://www.nngroup.com/articles/which-ux-research-methods/

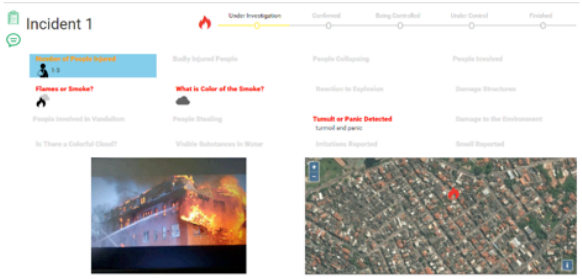
Appendix

EMERGENCY RESPONSE TOOLKIT

1. ERTK user sets the emergency management session with safety areas
2. ERTK user sets the crowd sensing session
3. ERTK shows the first incident alert

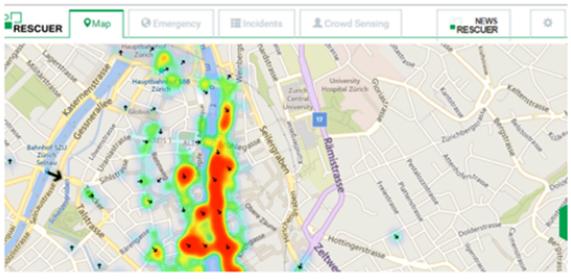


4. ERTK user visualizes the report information in the aggregation screen



5. ERTK shows aggregated incident information
(color of the information indicates the source of information)
6. ERTK user passes on an attribute and visualizes the image with its confirmation
7. ERTK user visualizes the word cloud (related to when, who and why)

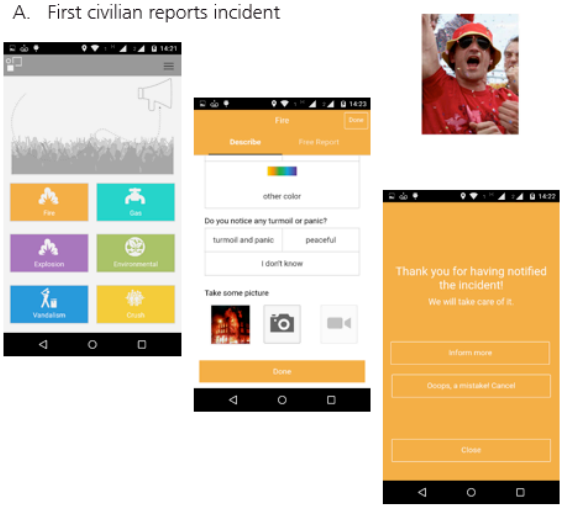
8. ERTK user decides to check the distribution of information regarding some non-visual attribute (injured people)
9. ERTK user visualizes the reports in the incident map (initially x Help reports, after the analysis only y, because z have been identified as Fire)
10. ERTK user visualizes the heat map



11. ERTK automatically changes incident status to "Being Controlled"

MOBILE CROWDSOURCING SOLUTION

A. First civilian reports incident

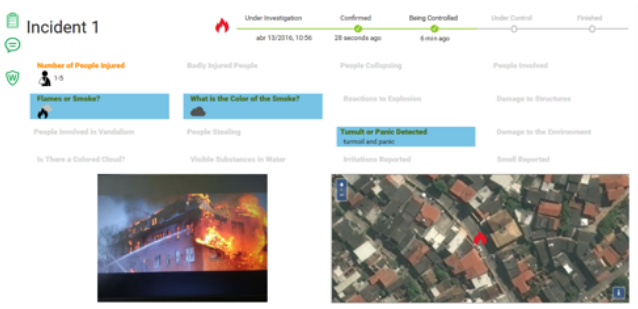


B. More civilians and supporting forces report the incident; x civilians report Help

C. Member of the workforce report incident – member of the workforce uses voice to say the streets that are affected by the smoke; voice is translated into text

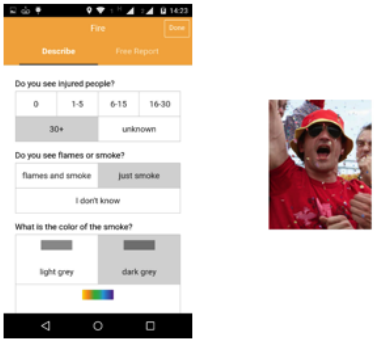
D. Further civilians, workforces keep reporting

12. ERTK shows current aggregated incident information (more precise due to the nature of the fields, e.g., size of the fire; the color of the attribute will indicate the information provided by the workforce)



13. ERTK user looks at on the information provided by the workforce about the incident, but do not try to confirm the information
14. ERTK user visualizes the heat map to check if civilians have followed the Guidance Message
15. ERTK user draws the incident zones
16. ERTK user sends Public Communication with incident status "Being Controlled"
17. ERTK user asks if the person is trapped or has movement issues

E. Civilian answers Information Request (person is trapped)



F. Member of the workforce answers Information Request with the status "Under Control"

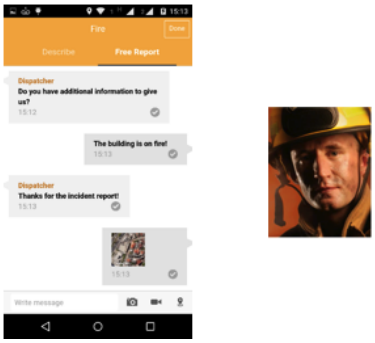


Fig. 12. Interaction between Emergency Response Toolkit (left) and RESCUER App (right)