Evaluating Novel User Interfaces in (Safety Critical) Railway Environments

Anselmo Stelzer, Isabel Schütz, and Andreas Oetting

Railway Engineering, TU Darmstadt, Germany
{lastname}@verkehr.tu-darmstadt.de

Abstract. In this paper we present a research facility for railway operations, the Eisenbahnbetriebsfeld Darmstadt (EBD) as simulation environment. Here, new operational and dispatching software for a safety critical environment can be thoroughly evaluated. In three expert and two user evaluations it could be shown that the EBD is a well-suited environment for testing as an extension to traditional methods. On the one hand, implementing software in the EBD can be done in a timely manner and at relatively low costs. Also, it is possible to trigger certain disruptions and malfunctions at will which would be impractical in real operations. On the other hand, studies have shown that users are really keen on testing in the EBD and that their mood is constantly good all along.

1 Introduction

The domain of railway operations is in most parts a safety critical environment with special requirements regarding used technology and software. The evaluation of the software used is more complicated than in standard environments, mainly because it cannot be tested and evaluated in the real environment. As for many other safety critical domains real testing environments exist, this is not possible for the actual railway operation. For a realistic evaluation an environment which comes as close to reality as possible is needed to produce realistic results and to make the assessors act realistically.

An evaluation and testing environment is crucial to show that developed software actually fulfils its mission. But tests in environments that simplify reality by simulation or abstraction are – depending on the level of abstraction – only partially transferrable to reality.

That is why an environment is needed for the railway domain to test and evaluate new software and interfaces which comes very close to real operations. In this paper we will present a possible solution for this problem.

It has been shown in different evaluations that, concerning mental effort, testing in this environment is very demanding for the user and can therefore be estimated as very realistic. Moreover, users tend to be really enthusiastic about testing in such an environment which increases the willingness for testing as such.

M. Kurosu (Ed.): Human-Computer Interaction, Part III, HCII 2014, LNCS 8512, pp. 502–512, 2014.

In section 2 we talk about the problems and our motivation for evaluation in railway operations and in section 3 we will present a possible solution for the problem, the Eisenbahnbetriebsfeld Darmstadt (EBD). We will present three practical uses in section 4 to show the suitability of our EBD. In the following chapter we will discuss the approach and conclude in section 6.

2 Related Work

Railway operation is a critical field of application where software has to work correctly and reliably. Thus, software of railway technology in the field of control, command and signaling is subject to very high requirements. Knight [1] points out problems that arise from safety-critical applications. Functional verification by testing is also difficult due to the environment this software is used in [1].

But the safety issue not only concerns the functionality and actions of a software in this domain, but also the dispatching actions a dispatcher decides to carry out based on displayed information. Today, almost everywhere in railway operations software is used to support railway staff in their work, mostly by displaying relevant information to staff to support them in their daily tasks. Some software provides functionality different from information display such as supervision of actions (in interlocking environments) or, very rarely, decision support [2,3,4].

Changing interfaces may entail seemingly missing or misleading information either through bad implementation, but also through staff not recognizing relevant elements in the new interface. As a result dispatching decisions cannot be carried out in the same quality as with the existing system. This is one of the reasons why introducing new interfaces in a railway operation system takes an unusual long amount of time.

Thimbleby [5] faces this problem by introducing Interaction Walkthrough (IW) for safety critical interactive systems. He explains that from an existing or prototype system another system is developed in which the remarks of the assessor are considered and implemented [5]. It is further argued that the changes achieved by IW are less expensive than programming from scratch and therefore significantly more economical [5]. For this approach to work out, the assessor has to be an expert to be able to discover discrepancies or malfunctions in the evaluated system.

Our aim goes a step further in the process of introducing new software. To make new or changed software being widely used, it needs to be accepted by its future users. Moore [6] distinguishes between continuous and discontinuous innovations. While continuous innovations are generally easily accepted by the vast majority of users, discontinuous innovations need to be carefully introduced. Moore defines a set of user types to classify the degree of innovations they can bear, thus there is a spectrum between continuous and discontinuous innovation [6]. Innovators and Early Adopters can cope better with discontinuous innovations early on while the (Early and Late) Majority need a certain amount of time to adapt to new technology and the Laggard needs an unusual high amount of time. The types are distributed in a bell curve (Figure 1).

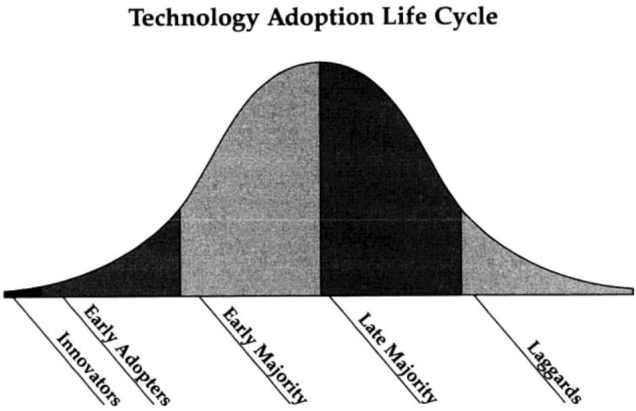

Fig. 1. Distribution of adopters for new technologies [6]

For the domain of railway operations we can assume a normal distribution for the railway staff within Moore's classification. That also means that the introduction of innovations is generally not easier than in standard user environments and the more discontinuous an innovation is, the harder for the staff to adapt.

An evaluation by an expert assessor will reveal malfunctions, badly designed interfaces and discrepancies in a system, but as an expert in the domain the assessor is rather to be classified as Early Adopter than a majority user. Accordingly, the degree of innovation he can cope with will be higher than that of the standard user. To evaluate the acceptance of average users an environment is needed where the interfaces can be evaluated without interfering with critical operations. Our aim is to create a setting in which new systems critical to railway operations can be evaluated most realistic with the future user as assessor. The evaluation is supposed to allow conclusions about the acceptance the usability of new interfaces.

3 Approach

The Eisenbahnbetriebsfeld Darmstadt (EBD) is a research facility for railway operations which embodies a very realistic simulation environment. It is operated by the Department of Railway Engineering, DB Training [7] and AKA Bahn [8]. In contrast to computer simulations the EBD comprises actual railway technology as far as possible. As a matter of fact only tracks and trains are models while the interlocking technology is real. Consisting of 13 stations, 160 main tracks and 380 points and derailers on about 90 kilometers of simulated line, the EBD offers a size in which scenarios of an adequate complexity can be created. Moreover, the EBD contains all generations of interlocking systems (mechanical and electro-mechanical signal boxes, relay interlocking systems and electronic

interlocking systems) and the respective dispatchers' work areas. Beyond infrastructural dispatching (actual railway operations), there are also work places for transportation dispatchers (amongst others rolling stock and staff dispatching, connection dispatching). The EBD is further equipped with technology that is generally used for railway operations such as phones with special dialing, walkie-talkies, and railway specific printed forms. This allows providing the complete chain of railway operation and dispatching to be mapped within the EBD. The EBD contains 13 computer work places (yellow rectangles in Figure 2) which can be switched between transportation dispatcher's and infrastructural dispatcher's work place [9].

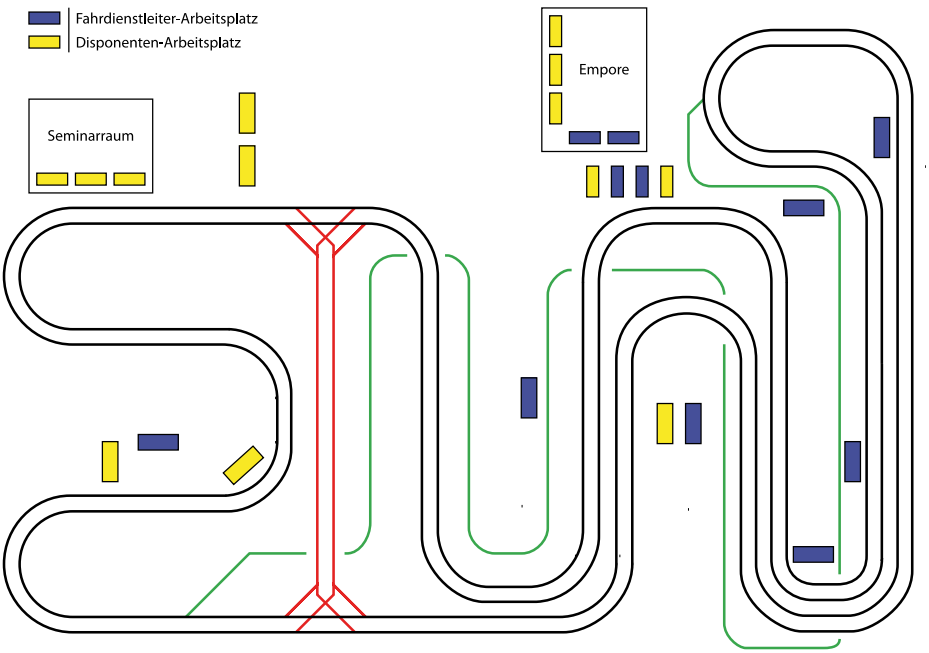

Fig. 2. Basic layout of infrastructure and work places in the EBD [9]

Beyond the working places there are seminar rooms. These are mainly used in seminars as training rooms, but they can also be used during the evaluation, i.e. for focus groups or interviews. The used software in the EBD is developed by one of the partners, AKA Bahn. With respect to the range of functions, the software corresponds to actual railway systems. Furthermore, implementation and user interface are identical to the real ones. This can be made sure on the one hand through the guidelines which were considered during the development of the software. On the other hand, among the members of AKA Bahn are traffic controllers, dispatchers, signalling technicians or employees of the Federal Railway Office who use their expert knowledge for this project. This creates an

independence from manufacturers, for example of interlocking technology such as Siemens or Thales [9].

All this makes the EBD as real as it gets and apart from that there are many advantages of testing in such an environment. If new elements are required for an evaluation they can be implemented in a timely manner. New interfaces can easily be attached and tried out, either in a standalone manner or alongside the existing interface, both in real operations mode.

Since the working areas are completely identical to the ones used in real railway operations with all the equipment needed, this makes sure that dispatchers and traffic controllers act in a realistic manner. It is not possible to cheat, i.e. by watching the trains or tracks; users need to fully rely on the software. This is achieved through the special construction of the EBD which has users sit away from the stations they are responsible for. Furthermore, besides realistic surroundings, it is possible to create disruptions and malfunctions in the operation at any time and in any manner, which is not possible in real operations [9]. This facilitates triggering relevant situations as needed for the evaluation. Since testers have no line of sight to the test leader, there are no indications for an upcoming event.

All this makes the EBD a predestined area for the evaluation of new interfaces and functions in the domain of railway operations.

4 Practical Evaluation

Recently, the EBD was used to evaluate several new visual approaches and modified interfaces in railway operations. Partly, the interfaces were developed at the Department of Railway Engineering of TU Darmstadt, partly cooperation was formed to evaluate externally developed interfaces. In both cases the Department of Railway Engineering as one of the operators of EBD took part in the evaluation process and is technically supported by AKA Bahn.

Both infrastructural and transportation dispatching support software have been evaluated in the EBD. For this it is generally necessary to create operational scenarios in which the tests can be performed; however, often it is possible to reuse existing scenarios and/or to extend them. This keeps the functional effort relatively small. Technically, the new software solutions need to be interfaced with the EBD. Again, the effort is very small compared to real environments, as existing software is developed in-house, therefore extendible and accessible via open interfaces.

Subsequently, we will present three approaches of interface evaluation in the EBD. Firstly, we will talk about an expert evaluation of a new way of displaying connection conflicts to the dispatcher. Secondly, we will focus on two user evaluations of a newly developed interlocking user interface and about a redesigned interlocking user interface. Both system and evaluation methods will be shown in detail.

4.1 Evaluating an Alternative Visual Approach for Connection Dispatching

In a research project with Deutsche Bahn, a new way of displaying connections and connection conflicts was developed. Part of the project was an evaluation of the new display on real conditions.

For the visualization each connection is assigned to a category. The way of displaying a connection is based on the same. The category advices the user how to proceed with a connection (conflict). The conflicts are arranged in a matrix depending on feeders and distributors (Figure 3). The alignment of connections in the matrix format is novel compared to the interface the transportation dispatcher is used to nowadays. Connections are presented in the cells of the matrix in case a connection for the feeder and the distributor exist. The background color of the cell depends on the category. Within the cell, additional information concerning the connection is displayed [4].

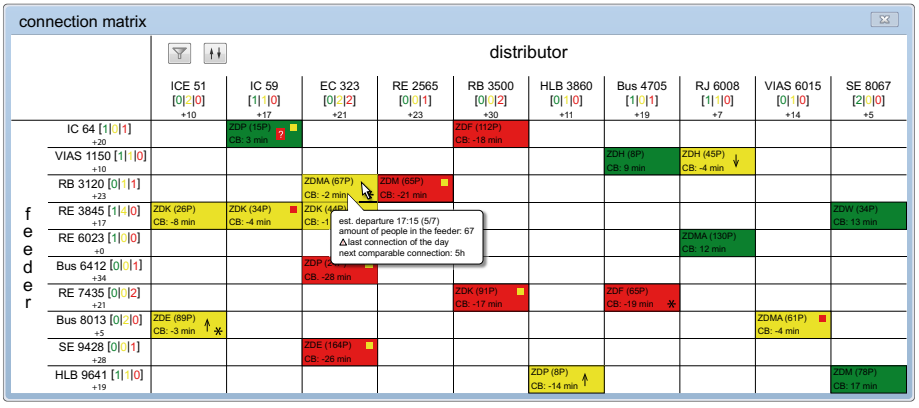

Fig. 3. Connection matrix [4]

Before implementing a prototype, several focus groups were hosted to gain input for the interface. The focus groups consisted of dispatchers which would have to work with the newly developed interface in the productive environment.

The EBD was used to proof that dispatchers are able to work with this new environment during a testing scenario with prospective users. The evaluation group consisted of dispatchers that, in addition to their daily dispatching routine, also work as trainer for quality enhancement seminars in the EBD. Therefore route knowledge was present among the participants.

To perform the evaluation of the new interface, an existing scenario was adopted and extended with connections and interchangers such that the matrix was filled with connections and information to be displayed for each of them. While running the scenario, several trains were delayed individually to provoke connection conflicts. Also, the evaluation group could choose specific trains to

be delayed to try out functions for a special connection pair. Apart from existing elements such as a schedule, the scenario contained rolling stock, staff, special waiting time rules and interchange times on platforms. Using the EBD, the effect in the software if different approaches to interchange times are used could be shown.

Since an existing scenario could be used as a basis, the cost of interfacing the new interface with the EBD was comparatively low. The prototype was exclusively developed for this environment and directly interacts with existing dispatching software already available at the EBD.

The evaluation revealed some interesting results even though the opinion of the dispatchers had already been obtained within the focus groups. The evaluation within the realistic environment of the EBD showed several aspects to be improved in the interface before it could be used in real operations. These aspects have been adopted to produce a revised version of the interface.

4.2 Evaluating a Newly Developed Interlocking User Interface

In cooperation with an external partner, a newly designed interlocking user interface was tied in with the EBD for a first evaluation. The interface is a complete innovation by combining a dispatcher and a traffic controller user interface. It combines functions which are nowadays spread over multiple programs. Moreover, it contains an event triggered approach. Also, the design as such is completely new and is reminiscent of an infrastructure modelling program. Thus it is possible to zoom in from an overview of the whole country up to a specific infrastructural element. The look and the functionality are completely different from electronic interlocking systems used today. Implementing this system will cause a change in the way of working and of operation as well as an adaption of the guidelines.

While interfacing the system with the EBD, an expert evaluation in the form of a Guidelines Review was conducted. Some experts from AKA Bahn compared handling and functioning of the software with corresponding descriptions in existing guidelines for railway operations in Germany. This revealed not only bugs in the system, but also some real fundamental errors in functioning. During these expert evaluations, focus was not only on correctness of function, but also on integrating as many functions as needed to test the system in a user evaluation. Since the experts used the interface for normal railway operations missing features manifested straightforwardly. Therefore it was possible to implement these features before further user testing. Ensuring that all essential functions have been implemented and keeping the amount of errors small is important for users testing the system because immature software can be frustrating to handle and therefore can distort the results of the study. Since on one hand it is not trivial to find an adequate amount of suitable users and on the other hand testing is costly, it is crucial to ensure that all results can be used. Moreover, completing and fine-tuning functions early on can drastically reduce time pressure.

4.3 Evaluating a Redesigned Interlocking User Interface

A study comparable to the one from subsection 4.2, but more advanced was realized with a redesigned interlocking user interface. In this system, the existing interface was enhanced by a new window framework. This framework is supposed to be gradually extended by all tools a traffic controller is presently working with. So only the design of the interlocking interface was slightly modified compared to the electronic interlocking used today, while functionality and way of operations remain the same.

Before actual user evaluation, the same procedure as shown in subsection 4.2 was carried out to ensure success of the study. The aim of the user evaluation which took place after the expert evaluation assessed the usability of the user interface and the mental effort the traffic controller faces during testing. The evaluation was divided into three parts: firstly a free exploration, secondly the accomplishment of a set of tasks and thirdly a scenario the traffic controller has to handle with the prototype. Free exploration aimed at familiarizing the user with the system. Focus in the second part of the evaluation was on some basic functions like zooming, window handling and panning. The last part aimed at simulating common operations of the traffic controller with the system. The basis for this scenario was a timetable with duration of about one hour which was specially developed for this evaluation. It contains 34 trains per hour of different types, e.g. commuter and long distance trains, with different passenger stops, intervals and velocities. Through comparison with real timetables and through expert evaluations it could be ensured that the timetable is as close to reality as possible. Moreover, the scenario contains seven events within half an hour. These were chosen according to the frequency they occur in reality and with consideration of the implemented functions of the prototype. Examples for events are a defective train door, a point failure or alarm of the hot axle box detector.

At the beginning of the evaluation, users were requested to fill in a demographic questionnaire and a questionnaire assessing their current mood. After each session, users were asked to state their mental effort and their opinion about the system, also by using questionnaires. Since it was the first such user evaluation, it was aimed to keep equipment and therefore costs short, which is why the mental effort was assessed using questionnaires instead of physiological measurements. Eventually, the users were asked one more time to state their current mood by using a questionnaire and they were invited to take part in an interview to state their opinion about their experience in using the prototype. During the whole evaluation, a video camera was used. Moreover, the users were observed during the whole evaluation by the test leader. Additionally, users were encouraged to think aloud during the whole study.

On one hand, the study has revealed many facts and improvement proposals about the prototype, but it has also shown many interesting facts about testing in the EBD. It could clearly be observed that perceived mental effort measured by the "Subjective Mental Effort Questionnaire" (SMEQ) increased from "hardly demanding" after the set of tasks to "quite demanding" and even

"strong demanding" after the scenario. This is the result of the realistic and demanding scenario with many events. Moreover, it was revealed that users had no positive impression of the prototype. The opinion of the users about the prototype substantially worsened during the evaluation and, as a result, enthusiasm declined drastically. Astonishingly, the mood of the users was constantly good. They felt as active, engaged, awake and attentive as before. During the interviews, it could be stated that participants rated the evaluation itself very positively. They were really keen on evaluating in the EBD and were fascinated of the infrastructure. One difficulty that manifested itself during the scenario was finding the location of events. Users had to face the difficulty that neither the prototype nor the locality where known to them. In case of an upcoming event, users had a hard time finding the respective location. Only after having located the origin, they could start dealing with the event. Then they had to look for the right software feature, so time for processing only one event was substantial. Since the time for the scenario was limited to half an hour, events were triggered in short succession. That's why during evaluation, users had to cope with many events in parallel because most of the time, events could not be processed completely before the next started. This was very demanding and sometimes confusing for the users.

5 Discussion

The EBD offers great opportunities to evaluate new functions and interfaces in a real environment. Whereas the environment is as close to reality as a simulation can be, interfacing costs are very small compared to an implementation in real operations.

Still, implementing a new interface and a scenario in the EBD is a substantial effort, as the scenarios have to be developed and new components need to be interfaced with the EBD. That is why different methods should be applied beforehand. This includes for example an Expert Evaluation, Focus Groups as performed in subsection 4.1, Guidelines Reviews as performed in subsection 4.2 and subsection 4.3 or the Interaction Walkthrough proposed by Thimbleby [5]. Heuristic evaluation as [10] can also be envisioned.

After having ensured that the evaluator faces an application that has already a certain degree of quality and a sufficient range of functions, it can be tested for practical application by common users. The comparably easy integration of new technologies also enables several evaluation sessions with an improved version of the new interface after feedback from the evaluator.

There is a broad spectrum of methods which can be used to evaluate a prototype with users. In subsection 4.3 we have performed an evaluation using video recording, questionnaires, observation and interviews. Users were highly motivated, enthusiastic und dedicated, even after having tested for about two hours, not being delighted about the prototype and facing an increasing mental effort. Larger scale studies can be conceived which facilitate psychophysiological measurements, eye- or mouse-tracking to measure mental effort much more precisely. Using these evaluation methods, it must be considered that having no

route knowledge can cause the evaluator to have an increased mental effort and to look around in a confused way. For this it might be optimal to develop some sort of filter to discern glances which are looking for the right location along the track, but this might not be superficial.

One disadvantage of evaluating in the EBD is the lack of route knowledge. Finding the right location in case of occurring events may take some time and can distort the measurement of reaction or processing times. Therefore, timing processing or reaction durations is problematic because results can be misleading. In general, users in field studies need more time to perform tasks as it was observed by [11]. In reality, traffic controllers have to take route knowledge examinations and are not allowed to work independently before having passed [12]. This fact cannot be considered during testing because it is too time-consuming to teach the users before testing. Moreover, studies have shown that testing is very exhausting for the users. So the duration of tests must be limited to one or one and a half hour.

Although it is very easy to implement some form of logging every action the user executes on the prototype, results would as well be distorted because of missing route knowledge. Maybe after having conducted additional evaluations or seminars in the EBD, as for the assessors in subsection 4.1, it might be possible that more of this data be valid because the users accumulate route knowledge. Missing route knowledge should also be considered during creation of the scenario. There should be enough time between the events so that users do not have to cope with two events at the same time. This might also distort the results because mental effort increases drastically.

Summarizing, it can be stated that the EBD is a very good environment for usability tests as an extension to traditional methods which helps to achieve feedback about the quality and usability of newly developed software or interfaces.

6 Conclusion

First evaluations have revealed the possibilities the EBD offers for evaluations. Therefore much more research should be done. Fields of investigation are for example the development of suitable scenarios. It is necessary to define the adequate number of events to be scheduled within a certain amount of time. Also, some larger scale studies in the field of mental effort using psychophysiological instruments might be useful. This may confirm results obtained from a first study comprising questionnaires. A more general objective might be a comparison of the results attained in field studies, laboratory studies and simulation. Probably, a comparison with the kind of simulation being performed in the field of aviation may reveal interesting facts.

Based on the results of the evaluation presented in subsection 4.1 an implementation in a real control center is carried as the next step for this project. Thus, the EBD can help to gain knowledge about the usability of newly designed interfaces and their potential for the real environment.

References

1. Knight, J.C.: Safety critical systems: challenges and directions. In: Proceedings of the 24th International Conference on Software Engineering, ICSE 2002, pp. 547–550. IEEE (2002)
2. Fay, A.: Wissensbasierte Entscheidungsunterstutzung fur die Disposition im Schienenverkehr: Eine Anwendung von Fuzzy-Petrinetzen. VDI Verlag, Düsseldorf (1999)
3. Kurby, S.: Makroskopisches Echtzeitdispositionsmodell zur Lösung von Anschlusskonflikten im Eisenbahnbetrieb. PhD thesis, Technische Universität Dresden, Dresden (February 6, 2012)
4. Stelzer, A., Oetting, A., Chu, F.: Connection Dispatching - an Algorithmic and Visual Support for the Dispatcher. In: Proceedings WCTR 2013 (2013)
5. Thimbleby, H.: Interaction walkthrough: Evaluation of safety critical interactive systems. In: Doherty, G., Blandford, A. (eds.) DSVIS 2006. LNCS, vol. 4323, pp. 52–66. Springer, Heidelberg (2007)
6. Moore, G.A.: Crossing the chasm: Marketing and selling technology products to mainstream customers, Rev. ed., [nachdr.] edn. HarperCollins Publ., New York (2005)
7. Deutschen Bahn AG: Website of DB Training Learning & Consulting (2014)
8. AKA Bahn e. V.: Website of Akademischer Arbeitskreis Schienenverkehr e.V. (2014)
9. Streitzig, C., Stelzer, A., Schön, S., Chu, F.: TU Darmstadt – Research, Training & More Besides. EURAILmag, 152–159 (2012)
10. Nielsen, J.: Heuristic evaluation. Usability Inspection Methods 17, 25–62 (1994)
11. Duh, H.B.L., Tan, G.C., Chen, V.H.H.: Usability evaluation for mobile device: a comparison of laboratory and field tests. In: Proceedings of the 8th Conference on Human-computer Interaction with Mobile Devices and Services, pp. 181–186. ACM (2006)
12. Hausmann, A., Enders, D.H.: Grundlagen des Bahnbetriebs. Bahn Fachverlag (2007)