Acta Informatica Pragensia, 2020, 9(2), 108–131
DOI: 10.18267/j.aip.136
Original Article

Research on User Behaviors and Tolerance of Faulty Web Interactions

Veli Özcan Budak [1], Emre Akadal [2], Sevinç Gülseçen [2]

Abstract

Even if we think that all the computer systems that are in operation work perfectly, the background might not be as it seems. We might face some faulty web interactions on a popular website or software as well. User behaviors are vital for developers in creating a satisfying computer system. in the aim of this study was threefold. Firstly, to determine if users' tolerance of different kinds of faulty web interactions changes depending on the environment, and then to find how users' behaviors differ when they encounter a faulty web interaction. Lastly, to detect how faulty web interactions shape users' perceptions. To achieve these aims, we conducted a test on a manipulated mobile e-commerce website with 11 tasks including five faulty ones. Participants were not informed that the test includes faulty tasks. Faulty tasks consist of different kinds of web errors: Not Responding, Blank Page, Connection Timeout (HTTP-500), Not Found (HTTP-404), and Redirect (HTTP-301). The other tasks were organized as dummy tasks, and they were not examined. In the results of this study, we reached quantitative (for the collection of quantitative data, we used a Tolerance Evaluation Scale (TES) that we developed for this study) and qualitative findings. According to the quantitative findings, there is no difference between the tolerance levels of users for different environments. On the other hand, it was determined that when there is an error that includes feedback, user tolerance is affected positively. In addition to this, it can be seen that users have a low tolerance towards giving another chance to any kind of website which has a faulty interaction. In terms of qualitative findings, participants emphasized that it does not matter what purpose a website serves, the errors give an amateur impression by damaging usability and professionalism.

Keywords: User behavior, User tolerance, Faulty interaction, Mobile web, Human-computer interaction, Usability.

1 Introduction

Because of the fact that making mistakes endlessly is in human nature (Norman, 2013), we are able to see different kinds of errors in every kind of human action (Phalgune et al., 2005). The cause of a mistake may consist of various factors (Begosso & Filgueiras, 2006). Detecting errors and recovering from them has been a research topic for a long time. System-based or user-based errors are observable in human-computer interaction (HCI), which is a research field with powerful methods such as eye-tracking, electroencephalography (EEG), interviews, etc. HCI

[1] Department of Information Technologies, Kirklareli University, Kayalı Kampüsü Merkez, Republic of Turkey

[2] Informatics Department, Istanbul University, Kalenderhane Mah. 16 Mart Şehitleri Cad. No: 8 PK 34134 Vezneciler-Beyazıt-Fatih/İstanbul, Republic of Turkey

✉ emre.akadal@istanbul.edu.tr

has two complicated entities trying to "speak" the same language. These entities are human beings and computer-like machines. This kind of communication comes with some difficulties.

Lazar, Meiselwitz, and Norcio (2004) stated that there are four different types of error when using the World Wide Web (WWW). These are user errors, system errors, situational errors and poor web design. User errors occur because of incorrect user actions, system errors are related to software or hardware problems, and situational ones are errors such as network errors. On the other hand, poor web design is related to websites designed in a confusing manner. Even though these authors grouped errors in four categories, it can be said that the four error types are connected to two different kinds of errors: user-based (user errors) and system-based (system and situational errors and poor web design). Similar to these error types which are stated by Lazar, Meiselwitz, and Norcio (2004), Ma and Tian (2007) also grouped web-based errors in 3 different categories: host, network and browser errors, source and content errors and user-based errors. Even if these group names are different from Lazar, Meiselwitz and Norcio's (2004), the errors are related to the same types. In other words, except the user-based errors, the first two types are among the error types that can occur independently from users. For this reason, they can also be defined as system-based errors.

As people who use a system/software have different backgrounds and knowledge levels (Graham, 2003), user-based errors might occur differently. Nevertheless, an error which is made by a user in using a computer system for an individual purpose would affect just the user who made it. In these kinds of situations, user guides included in systems/software can be one of the solutions for reducing the errors made by users. On the other hand, system-based errors, which occur independently from users, need to be approached differently, since these kinds of errors, which can be considered as the background functionality of a computer system, will affect all users of that system. When considering each scenario, it can be said that these two kinds of error occurrence do not have the same effect on users. Especially when we imagine that system-based errors might lead to catastrophic user experiences, this case might also result in expensive recovery processes (Heckel & Mariani, 2005). In the light of these explanations, even if user-based errors can be tolerated in the context of individual usage, it is not possible for the system-based errors. In other words, the fact that the final product will most likely contain some faulty interactions that might be user-based or system-based does not mean that system-based errors can be ignored by developers, since faults constitute a critical threat for the dependability of computer systems (Ploski et al., 2007). Meyers (2004) stated that "responsibility for interface usage errors belongs to the interface designer, not the interface user". Similar to Meyers' (2004) statement, we can easily indicate that responsibility for system-based errors always lies with the developers/designers.

A fault has been basically defined as a structural imperfection in a system that might end with unexpected results (Munson et al., 2017) or the cause of an error that means the delivery of a service is not performing as expected (Laprie, 1995). From the perspective of system-based errors, faults might occur for different kinds of reasons. It might be a functionality error unnoticed during the development process or a design error made by an inexperienced developer. Furthermore, a well-performing system might not work at another time in the – near or distant – future because of being updated. This situation can also be counted as an error type. Even if the behavior of making mistakes is in human nature and cannot be avoided in terms of user-based errors, every kind of system-based errors can be fixed by a direct intervention of developers. Developers' actions can reduce or eliminate system-based errors completely. The most usual action for this is to remove or repair the cause of the fault (Avizienis, 1978). No matter what kind of error it is, the main point which has to be considered is how system-based errors affect users and what kind of errors should be taken more seriously in order not to lose

the users in every aspect, since it should not be forgotten that the reliability of a computer system depends on understanding the impact of the faults (Ocariza et al., 2013).

Generally speaking, systems always have undetected errors, and users always make mistakes. When a user encounters a fault on a system, she/he can easily blame the system or just herself/himself. This endless loop can and will not be broken. However, looking from the perspective of system-based errors, this situation can also be considered a part of quality. The products which have a lot of faults are considered as having poor quality (Card, 1998). Rubin and Chisnell (2008) stated that "usability is a quality that many products possess, but many, many more lack". From this perspective, it is obvious that the system-based errors should be reduced to minimum levels by developers so that a system can be considered high quality or usable.

Usability, which was defined as "the effectiveness, efficiency and satisfaction with which specified users achieve specified goals in particular environments" by ISO (1998), has five fundamental components: learnability, effectiveness, memorability, satisfaction, and errors (Nielsen, 1993; Shneiderman & Plaisant, 2005). There is a lot of studies concerning each component, but in this study, we focused on errors. In other words, how user behavior, tolerance, and perception change when encountering faulty interactions. Some studies which might be considered related to ours and contain user behaviors, are mentioned below:

Ramsey, Barbesi, and Preece (1998) performed a study by injecting delays into the page loading process. Their aim was to examine whether the latency between requesting a page and receiving it affects user perceptions. In this research, they found that faster pages are more interesting than slower ones. In addition to this, being slow results in a reduction of user motivation and increases user frustration. Tzeng (2004) carried out a research aiming to understand how users react to computers' apologies. In this study, in which a computer guessing game was designed, some minor flaws were intentionally integrated into the game, such as repetitively selecting the same keys and clues, an unattractive interface, irrelevant clues. The aim of this integration was to create a reason for the computers to apologize. This action can be considered a manipulation similar to that of our study. The results of this study show that even if some subjects felt manipulated when the computers offered apologies to them, the computer apologies helped to create more desirable psychological experiences for the users. Another study that examined the effects of different delays on two websiteswas carried out by Galletta et al. (2004). The authors created two manipulated websites in order to observe user behavior in a total of 196 participants. The results of this study show that an increase in delay time(s) affects performance, attitudes, and behavioral intentions negatively.

Another study was conducted by Everard and Galletta (2005), aiming to explore whether website presentation flaws affect consumers' perceived quality of the online store, trust and consumers' intention to purchase from the online store. They used three types of manipulative factors: a poor style (contrast and design flaws), incompleteness (placeholders such as "under construction" or "image not yet available" on each page) and language error (making a grammatical error on every page). The results of this study show that every kind of flaw that was tested affects users' perceived site quality, trust, and users' intention to purchase negatively. Guse et al. (2015) conducted a study assessing how delayed loading and partly loading webpages affected users' perceived quality. The authors of this study, which focused on Task Completion Time (TCT) and Page Load Time (PLT), concluded that PLT and TCT alone are not sufficient quality indicators when considering partial load failures. Another research on exploring the relationship between response time and user perception in the context of smartphone interactions was conducted by manipulating the response times for four tasks in three applications (Tan et al., 2019). The authors of this study found that while switching

between pages, interfaces with a loading animation affect user tolerance positively. This loading animation can be understood as feedback, which this study will emphasize and also focus on the way in which this feedback is important for user tolerance.

In this study, in the light of the explanations and the studies mentioned above, we aimed to detect the differences in users' behaviors and perceptions and to investigate users' tolerance when they encounter a faulty web interaction on a manipulated mobile e-commerce website. Our research questions were as follows:

- How does users' tolerance of different kinds of faulty web interactions change depending on the environment?
- How do users' behaviors differ when they encounter a faulty web interaction?
- How do faulty web interactions shape users' perceptions?

In order to answer these questions, we created an e-commerce website which is specific to this study and includes some faulty interactions. Participants were requested to complete all tasks connected to a scenario. The scenario had 11 tasks, including five faulty ones. The remaining tasks were dummy tasks, which were used in order to convince users that faulty tasks were not integrated into the website intentionally and to secure the objectiveness of results. Findings include various metrics and indicators. The next section describes the method of the study. The third section shows our findings with various metrics and indicators. In the fourth section, we present our conclusion.

2 Method

In this study, the usability test method, which is used to determine the weaknesses of any product, was used differently. Instead of detecting weaknesses, it was used to examine user behavior, tolerance and perception on a mobile website containing various intentionally placed faulty web interactions. Before conducting the test, we built a mobile compatible e-commerce website and placed five kinds of errors appearing as system-based errors. After that, we planned a scenario which consisted of 11 tasks including five faulty ones. Working tasks (dummy task) were placed in order to distract the participants' attention so that they would not realize the faulty tasks had been placed intentionally. The details of the method of the study are described in subsections.

2.1 Digital Test Environment

The website was built as an e-commerce website that included various products such as computers, mobile phones, software, etc. Even though the website did not include any sale or payment process, it was designed as if it had such components. We introduced the website to the participants as newly built and as if it were to be put into service soon.

The design of the website was in a responsive structure in order for it to be compatible with mobile browsers. Thus, the study could be performed on the participants' own mobile devices. A view of the website in a mobile browser is presented in Figure 1.

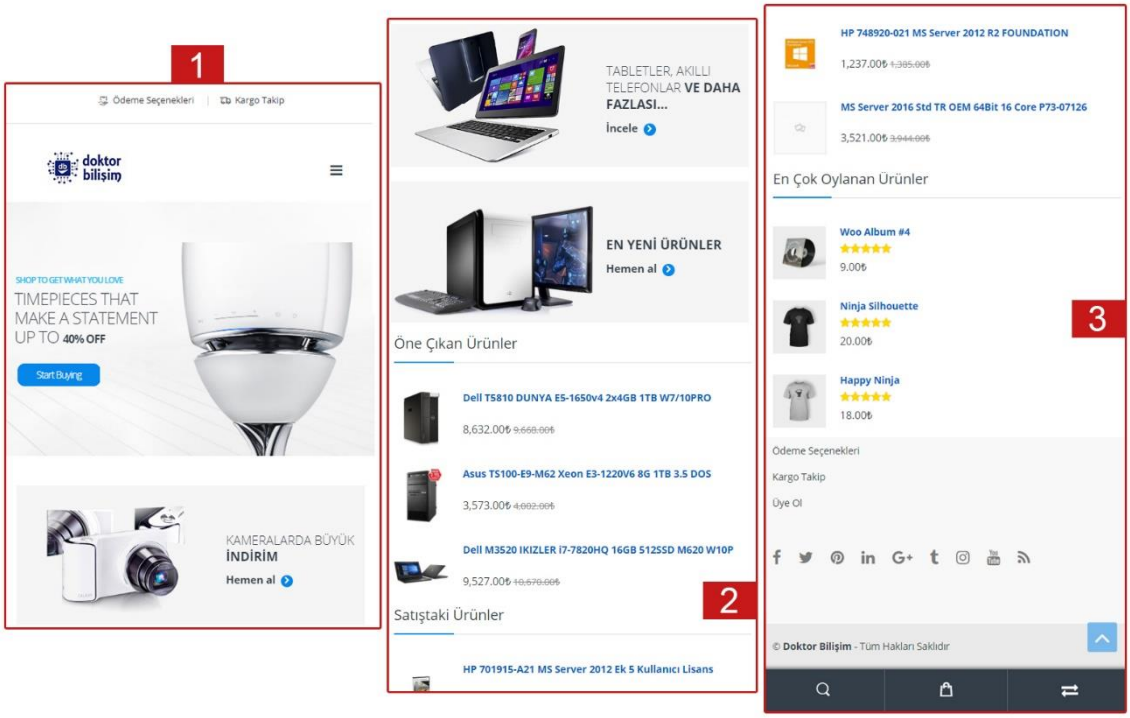


Figure 1. *View of the Website in a Mobile Browser. Source: Authors.*

The view of the main page of the website was divided into three parts because of the height of the page.

2.2 Participants and Test Environments

Budiu (2014) stated that it is important to choose participants who have used their phone for at least three months. Therefore, we checked this information first in order to identify suitable participants. Then, we focused on choosing the participants who experienced internet shopping in the past. After checking all the volunteers, we decided on 14 graduate participants (including 3 females and 11 males) who were appropriate. Each participant signed a consent document before the individual sessions started. After the selection process, the participants were divided into two groups, and the sessions were held in different places. The first place was a room that was customized for the sessions in Kırklareli University Distance Learning Implementation and Research Center for the first nine participants. The other places were the participants' own houses. This type of test environment was stated as "Informal Lab" by Barnum (2010). The reason for conducting home sessions is to investigate how participants behave in their natural environments.

2.3 Test Scenario and Tasks

In order to keep participants' motivation high, we prepared a scenario. In this way, participants had a single and exact purpose instead of independent tasks, as suggested by Barnum (2010). The introduction scenario was as follows.

"You got a job for the first time and you're waiting for the salary day, excitedly. You said to a friend of yours that you are looking for a budget-friendly e-commerce website for technological shopping. Your friend recommended a website which he/she did shopping on before. Finally, you got the salary and visited the website recommended by your friend!"

After the introduction, the scenario continues with the tasks in Table 1, respectively.

Table 1. *Tasks in the scenario.*

Order	Task	Task Type	Error Type
1	You are curious about the payment options. Please, open the payment options page to find one that is suitable for you.	Dummy Task	
2	You found the best payment option for you! Firstly, you want to change your old notebook. Find the notebook list on the website.	Dummy Task	
3	You are satisfied with the prices and want to order one. But the website forces you to register. Please register with this credential on the website. E-mail address: example@xmail.com Password: A1234B	**Faulty**	Not Response (NR)
4	You looked at the notebook choices and picked one: "Asus™ ROG GL753VE-GC095T". Please add it to your shopping cart.	Dummy Task	
5	You remember that you need a second monitor in your house. You want to buy it as well. Please find the "Asus VS197DE" model monitor.	**Faulty**	Blank Page (BP)
6	The toner in your printer is almost empty. It can use only black toner for printing. Please find the toner list and sort it by ascending order.	Dummy Task	
7	Your antivirus license is about to end. Please access the software list to buy a new one.	**Faulty**	Connection Timeout – 500 (CT)
8	You have suffered a power cut in your region. You worry that it can affect your new devices. Because of that, you want to look for Uninterruptible Power Supply (UPS). Please find the list.	Dummy Task	
9	You have chosen a UPS, "Dexter™ 850VA." Please open the detail page of this product.	**Faulty**	Not Found – 404 (NF)
10	Your brother wants a desktop computer and asks you to suggest one. While browsing, you saw "HP™ Z240" in the "new arrivals" section. Please find the stock code of the product to give him.	Dummy Task	
11	You are ready to finish your shopping! You want to check what you added to your shopping cart. Please open your shopping cart page.	**Faulty**	Redirect – 301 (RE)

The "faulty tasks" were taken into account for this study. However, we added "dummy tasks" in order not to lose the motivation of the participants. This type of task is not examined, but it is believed that this is essential to get valid and realistic findings. If the participants realized that the faulty interfaces are fiction, their behaviors could get unrealistic. It was determined at the post-test interview that no participant perceived that faulty tasks were placed intentionally.

In task 3, the participants faced a webpage which did not respond to clicking on the "register" link. In the regular process, they would have seen the registration form normally. Figure 2 shows screenshots from task 3.

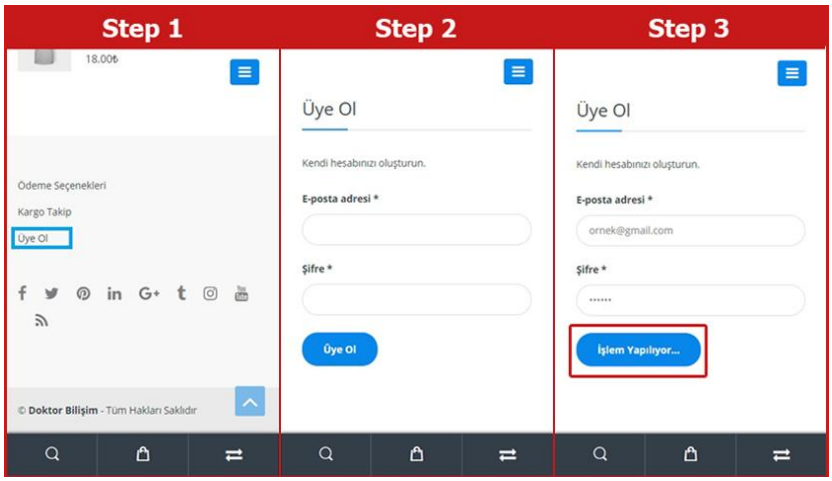


Figure 2. *Screenshots of Task 3. Source: Authors.*

In the faulty interaction, when the signup button is clicked, the label of the button changes to "processing". However, the process does not continue.

In task 5, we asked the participants to find a product on the website. They tried to access the product page in various ways (interaction 1: reaching the product list from menu or interaction 2: search box, Figure 3). We removed shortcuts that help to add any product to the shopping cart simply in the product lists. Participants were forced to access the product page. Even if they followed the correct way to reach the related product, they were not able to complete the task since the product page had been replaced with a blank page. In other words, they could not add the product to their shopping cart. Figure 3 shows a view of the product list.

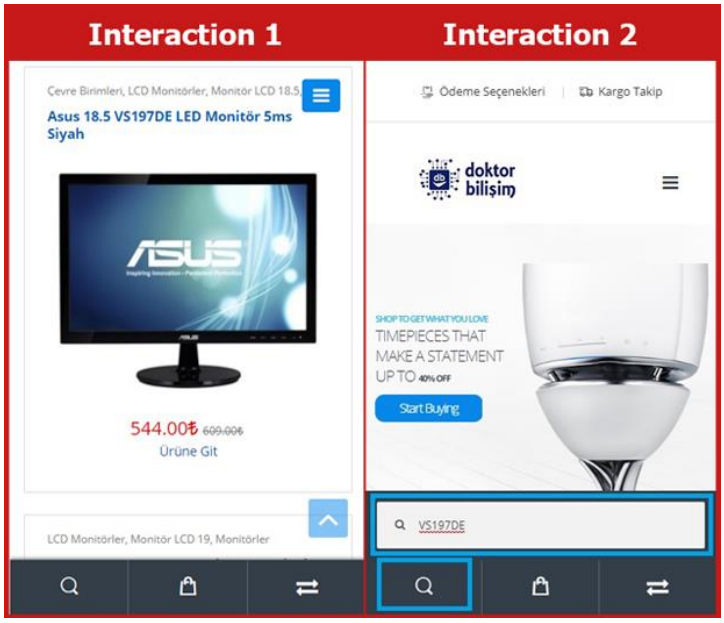


Figure 3. *Screenshots of Task 5. Source: Authors.*

In task 7, we asked the participants to open the software list to find an antivirus software. This faulty interaction was triggered by clicking the "Software" item on the menu. The category list was adjusted to give the "timeout" error after waiting 10 seconds. The view of the error page is in Figure 4.

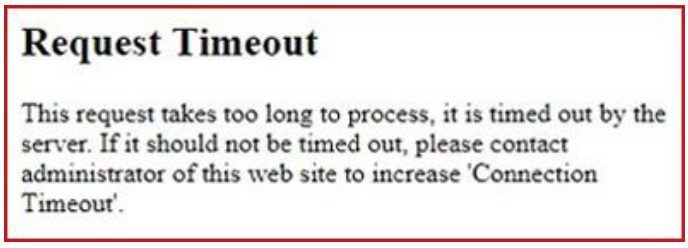


Figure 4. *Screenshot of Task 7. Source: Authors.*

Timeout duration was defined as 10 seconds according to the proposal of Nielsen (1993) who stated that the loading time of the page should be around 10 seconds in order to avoid the distraction of a user. Even though the newest studies in the literature suggest shorter duration, this study was predicated on Nielsen's statement.

In task 9, similar to task 5, we asked the participants to access the details page of the product. However, the link of the product redirected the participants to the NF error page as in Figure 5.



Figure 5. *Screenshot of Task 9. Source: Authors.*

In task 11, we asked the participants to access their shopping carts. But the shopping cart button redirected the user to the main page in every trial. Figure 6 shows the placement of the button.

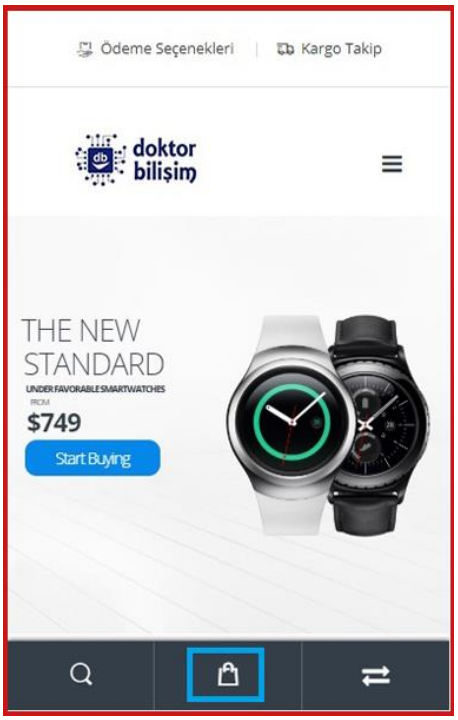


Figure 6. *Screenshot of Task 11. Source: Authors.*

In general, participants tried various ways to trigger all errors. However, the mobile website was manipulated in such a way that it was not possible to avoid encountering the mentioned errors. Thus, the participants were forced to interact with the errors.

2.4 Data Collection

In the data collection process, both qualitative and quantitative methods were used. While the qualitative findings include opinion and reaction of the participants during the process/end of test sessions, the quantitative findings were collected by the Tolerance Evaluation Scale (TES) that has been created for this research. TES consists of different kinds of metrics. The definitions and the descriptions of TES are given in Table 2.

Table 2. *Definitions and descriptions for TES.*

Definition	Abbr.	Description	Variable	Formulation
Starting to Interact	STI	The timestamp of the first user interaction with the website after she/he read the task.	t_1	t_1
Error Time	ET	The timestamp of the first triggering moment of the error	t_2	t_2
Error Duration	ED	Elapsed time from STI to ET.	t_3	$t_2 - t_1$
First Reaction Time	FRT	The timestamp of the first user reaction aiming to recover from the faulty interaction.	t_4	t_4
First Reaction Duration	FRD	Elapsed time from ET to FRT.	t_5	$t_4 - t_2$
The Number of Retries	NOR	The number of attempts to recover from the errors. The types of the attempts that we considered are as follows: n_1: Number of trying the search bar n_2: The number of refreshing the page	n_i	$\sum_{i=1}^{6} n_i$

		n_3: The number of turning back to the previous page and trying again n_4: The number of retries without refreshing the page n_5: The number of retries of the same pattern n_6: The number of trying a different pattern in order to complete the task		
Finish Time	FT	Timestamp of the most recent retry which has been finished.	t_6	t_6
Retrying Duration	RTD	Elapsed time in the retry process. It is also the difference between FT and FRT.	t_7	$t_6 - t_4$
Realize Time	RT	The timestamp of the realization that there is an error.	t_8	t_8
Realization Duration	RD	The elapsed time of the realization.	t_9	$t_8 - t_2$
Give-Up Time	GUT	The timestamp of the giving-up time of the participant to complete the task.	t_{10}	t_{10}
Give-Up Duration	GUD	The duration from the first triggered moment of an error to giving-up time	t_{11}	$t_{10} - t_2$

TES consists of 12 metrics to measure user behavior for the case of encountering a faulty interaction. Seven of the metrics (STI, ET, FRT, NOR, FT, RT, and GUT) are collected via observations during sessions. Among these metrics, RT and GUT are determined through the participants' verbal declarations (as in the sessions for this research). On the other hand, although NOR was calculated as described in Table 2 for this research, it can be calculated in different ways for different researches. For this research, in the calculation of NOR, we used six different user actions per error, but all types of user action did not occur in every instance. For example, n_4 was only calculated in the NR error, because the participants clicked the submit button again and again without refreshing the page. To sum up, NOR is a flexible variable that can be changed depending on user actions for every individual research. The remaining five metrics of TES are calculated by using the collected seven metrics. We defined the abbreviations for the metrics so that they could be easily used in the text.

In case of a faulty interaction, TES can be used to inspect the process in terms of determining the effects of an error on users. Thus, user behaviors can be foreseen in any kind of faulty interaction that can occur in any kind of system. The values collected by TES might differ for a website, a mobile application or user type. Users of a banking website or of a news website would probably not act in the same way. Consequently, TES can be used for different situations to measure users' tolerance of any kind of system errors. We believe that TES can be developed for different types of errors in future works as well.

We created a form in order to obtain quantitative data about the TES variables. This form was filled in both during the sessions and by watching video records after the sessions were completed. Thanks to TES, we were able to observe the participants while they were struggling with the mentioned errors.

2.5 Test Process

In this study, face-to-face interviews were first conducted with the volunteers. In these interviews, the volunteers filled in a form that included demographic questions. After the

participant selection process, the day and time when they would participate in the study was agreed. On the test days, the participants signed a consent document and read the introduction of the scenario. Before starting the sessions, all participants were asked to turn their mobile phone to airplane mode. In the sessions that were held in the university environment, Wi-Fi connection provided by the university was used to access the website. Similarly, home Wi-Fi connections were used in home environments. After the participants stated that they were ready, the sessions were started by the moderator. After completing all tasks and conducting the interviews, the sessions were ended.

In order to test whether the faulty tasks worked well, we performed a pilot test with the first two participants. After some corrections were made, the real tests were performed with the other 12 participants. The results of the pilot test were removed from the findings. We did not give any rewards to the participants in order to support their motivation; we thanked them instead.

2.6 Data Analysis

In the data analysis process, we examined the TES findings, voice, and video records. While qualitative findings were the voice and video records, quantitative findings consisted of the TES variables. In order to answer our third research question, qualitative findings were clustered by similarity and later discussed. For quantitative findings, firstly, descriptive statistics such as mean, standard deviation, etc. were used. After the explanations of descriptive statistics, we discuss significant test results based on our first two research questions.

2.7 Limitation

Problems due to the mobile device, server, or internet infrastructure are the limitations of this study. Also, page loading duration is different for various devices, at various times. However, our controls on these factors demonstrated that there was not any negative effect on the study.

3 Findings

The findings of the study are divided into two sections as quantitative and qualitative findings. The results of dummy tasks are not given as findings.

3.1 Quantitative Findings

In this section, we describe the findings collected by TES into 6 groups: ED, FRD, NOR, RTD, RD, and GUD, which are indicated in Table 2. The duration information in these variables is given in seconds (the NOR variable does not include any time data). Environment 1 (E1) indicates the test room in Kırklareli University Distance Learning Implementation and Research Center; Environment 2 (E2) indicates participants' own houses. All Environments (AE) represents both E1 and E2. The findings for task 3 are given in Table 3.

Table 3. *Findings for task 3.*

Environments	Participants	ED	FRD	NOR	RTD	RD	GUD
E1	1	78	20	3	104	47	137
	2	43	0	0	0	72	88
	3	133	41	2	2	79	84
	4	67	42	4	99	144	153
	5	45	151	1	34	215	215

	6	38	0	0	0	60	69
	7	67	122	4	50	212	221
E2	8	33	15	2	148	190	215
	9	83	207	2	3	214	227
	10	57	43	7	194	251	256
	11	52	32	4	54	53	117
	12	49	0	0	0	11	34
E1	**M**	67.29	53.71	2	41.29	118.43	138.14
E2	**M**	54.8	59.4	3	79.8	143.8	169.8
AE	**M**	62.08	56.08	2.42	57.33	129.00	151.33
E1	**SD**	32.60	59.63	1.73	45.36	71.84	62.21
E2	**SD**	18.14	84.12	2.65	87.50	105.40	92.16
AE	**SD**	27.22	67.24	2.11	65.57	83.81	73.93

When AE is considered, the mean of ED was found to be 62.08 seconds (SD: 27.22). It was determined that after encountering this error, the participants behaved patiently (FRD: 56.08, SD: 67.24). The case of zero "0" valued FRD, RC, and RTD shows that the participants did not take any action after they had encountered the error. FRD of the participants ranges between 15 seconds and 207 seconds. It can also be seen that the participants tried to recover from the error 2.42 times on average for AE (SD: 2.11). As the number of our observations was small (n=12), we used Spearman's ρ (rho) test in order to find association between the TES variables for this error. The results show that the GUD variable has a strong positive relationship with FRD (r_s=.787; p<.05), NOR (r_s=.63; p<.05), RTD (r_s=.674; p<.05), and RD (r_s=.851; p<.05) as expected. Additionally, it can also be seen that FRD has a strong positive relationship with RD (r_s=.775; p<.05) and NOR has also a strong positive relationship with RTD (r_s=.813; p<.05).

In task 5, users were faced with a blank page when trying to access a product page. The related findings are given in Table 4.

Table 4. *Findings for task 5.*

Environments	Participants	ED	FRD	NOR	RTD	RD	GUD
E1	1	24	10	7	43	25	53
	2	28	18	1	0	35	41
	3	29	11	3	137	140	153
	4	27	5	8	181	186	193
	5	18	12	5	106	122	126
	6	22	11	5	102	98	116
	7	30	6	9	236	226	248
E2	8	21	10	7	25	31	37
	9	26	9	5	165	168	180
	10	107	10	12	104	20	121
	11	17	12	9	133	120	150
	12	28	11	5	122	34	136
E1	**M**	25.43	10.43	5.43	115.00	118.86	132.86

E2	**M**	39.8	10.4	8	109.8	74.6	124.8
AE	**M**	31.42	10.42	6.33	112.83	100.42	129.50
E1	**SD**	4.31	4.28	2.82	79.73	73.85	73.52
E2	**SD**	37.81	1.14	2.97	52.34	65.79	53.69
AE	**SD**	24.18	3.23	2.96	66.86	71.19	63.35

The mean elapsed time to face the error for the first time (ED) was 31.42 for AE (SD: 24.18). It was determined that according to the FRD, FRT, RD, and the GUD variables, the participants behaved more impatiently in this task, compared to task 3. At the same time, the participants realized there was an error faster than in task 3. They interacted with the error after 10.42 seconds on average for AE (SD: 3.23). For the NOR variable, low standard deviation values show that the participants' behaviors are similar to each other. They retried to recover from the error 6.33 times on average (SD: 2.96), and this took 112.83 seconds (SD: 66.86) for AE. The GUD variable shows that participants could tolerate this error for 129.5 seconds on average for AE (SD: 63.35). In addition to these findings, it can be seen that task 5 is the most retried task by the participants. Since there was not any feedback or sign about what was happening, this error type seemed confusing to the participants. According to the findings from Spearman's ρ test, while GUD has a strong positive relationship with RTD (r_s=.993; $p<.05$) and RD (r_s=.832; $p<.05$), RTD has also a strong positive relationship with RD (r_s=.818; $p<.05$).

In task 7, the participants were made to wait for 10 seconds deliberatively after they clicked the related menu link, and then the CT error page was shown to them. In the findings of this task, some of the participants preferred to be patient until they saw the page while the others acted in the opposite way. The detailed findings are given in Table 5.

Table 5. *Findings for Task 7.*

Environments	Participants	ED	FRD	NOR	RTD	RD	GUD
E1	1	3	8	8	89	55	120
	2	4	12	5	45	32	81
	3	141	11	2	17	22	39
	4	2	3	9	123	56	149
	5	3	12	2	16	37	43
	6	3	14	1	4	29	29
	7	4	11	2	13	12	32
E2	8	3	16	2	23	43	73
	9	4	36	2	10	38	57
	10	3	14	2	18	47	47
	11	20	6	4	42	19	53
	12	3	17	2	38	46	64
E1	**M**	22.86	10.14	4.14	43.86	34.71	70.43
E2	**M**	6.6	17.8	2	26.2	38.6	58.8
AE	**M**	16.08	13.33	3.42	36.50	36.33	65.58
E1	**SD**	52.10	3.63	3.24	45.35	16.27	47.74

E2	**SD**	7.50	11.05	0.89	13.50	11.50	10.06
AE	**SD**	39.64	8.19	2.61	35.65	14.02	36.28

According to FRD, participants numbered as 1, 4, and 11 took action before seeing the error page. The mean of FRD was determined as 13.33 seconds (SD: 8.19) for AE. At the same time, NOR to recover from the error was 3.42 (SD: 2.61) in 36.33 seconds (SD: 14.02) for AE. Overall, although E2 participants both understood the existence of the error late (RD: 38.6) and acted more tolerantly at first interaction (FRD: 17.8) than E1 participants, they made less effort (NOR: 2) and gave up more quickly on recovering from the error (GUD: 58.8). Spearman's ρ test results show that FRD and NOR have a strong negative relationship (r_s= -.670; p<.05) as it is the same as between ED and RD (r_s= -.783; p<.05). These negative findings indicate that in the case when a user waits for any reason (we believe that even if there is not a faulty situation), the user tends to act impatiently and to give up easily. On the other hand, it can be seen that the GUD variable has a strong positive relationship with NOR (r_s=.803; p<.05), RTD (r_s=.846; p<.05), and RD (r_s=.712; p<.05). Additionally, NOR and RTD have a strong positive relationship (r_s=.897; p<.05), as expected.

In task 9, similar to the functionality in task 5, the participants faced the NF error page instead of seeing a blank page when they tried to access the product page. The findings are given in Table 6.

Table 6. *Findings for Task 9.*

Environments	Participants	ED	FRD	NOR	RTD	RD	GUD
E1	1	24	12	1	3	6	21
	2	22	0	0	0	9	12
	3	23	9	3	94	100	111
	4	19	12	1	4	20	26
	5	24	14	2	16	33	43
	6	15	8	1	5	17	17
	7	26	9	1	5	5	20
E2	8	12	18	1	2	17	36
	9	16	14	1	5	10	28
	10	20	14	1	1	8	22
	11	12	11	1	11	8	28
	12	15	11	2	19	25	37
E1	**M**	21.86	9.14	1.29	18.14	27.14	35.71
E2	**M**	15	13.6	1	7.6	13.6	30.2
AE	**M**	19	11	1.25	13.75	21.50	33.42
E1	**SD**	3.72	4.56	0.95	33.82	33.57	34.62
E2	**SD**	3.32	2.88	0.45	7.47	7.37	6.26
AE	**SD**	4.90	4.43	0.75	25.95	26.14	26

In this task, even though E2 participants detected the error more quickly than E1's (ED), the first reaction of E1 participants was slower (FRD). In addition to this, E2 participants gave up

recovering from the error more quickly (GUD: 30.2) when compared to E1's. When considering both environments, the mean of FRD was determined as 11 seconds (SD: 4.43), and NOR was 1.25 times (SD: 0.75) in 33.82 seconds (SD: 25.95). Another remarkable finding about this task is that even though this and the fifth task have the same functionality, the NOR values of this task are very different from the task 5. This finding indicates that participants retried much more often when not receiving an informative response. When we compare the findings of all errors, it can be seen that the findings of this error have the lowest values in general. From this point, it can be seen that because of the fact that feedback was given, the participants knew what they had encountered and they spent less time on recovery. In addition to these findings, Spearman's ρ test results show that similar to task 7, the GUD variable has a strong positive relationship with NOR (r_s=.840; p<.05), RTD (r_s=.677; p<.05), and RD (r_s=.698; p<.05). On the other hand, NOR has also a strong positive relationship with RTD (r_s=.844; p<.05), and RD (r_s=.691; p<.05). The findings for task 11 are given in Table 7.

Table 7. *Findings for task 11.*

Environments	Participants	ED	FRD	NOR	RTD	RD	GUD
E1	1	2	13	7	57	64	80
	2	2	15	4	16	31	50
	3	4	14	4	66	114	118
	4	11	10	13	188	163	188
	5	5	13	3	44	19	72
	6	9	17	2	13	17	39
	7	3	9	8	174	28	193
E2	8	2	19	4	73	49	102
	9	2	10	2	20	30	30
	10	2	28	3	36	48	86
	11	2	11	5	69	21	87
	12	7	33	3	88	76	138
E1	**M**	5.14	13	5.86	79.71	62.29	105.71
E2	**M**	3	20.2	3	57.2	44.8	88.6
AE	**M**	4.25	16	4.83	70.33	55	98.58
E1	**SD**	3.53	2.77	3.80	72	56.12	63.10
E2	**SD**	2.24	10.18	1.14	28.15	21.14	38.93
AE	**SD**	3.14	7.46	3.16	57.01	44.29	52.92

In this task, E2 participants quickly realized that there was an error (RD: 44.8) and acted more impatiently to recover from the error (NOR: 3, RTD: 57.2 and GUD: 88.6). Additionally, the mean of FRD of the participants is 16 seconds (SD: 7.46) for AE. At the same time, the participants tried to recover from the error 4.83 times (SD: 3.16) in 70.33 seconds (SD: 57.01) for both environments. According to Spearman's ρ test results, as in the GUD variable has a strong positive relationship with NOR (r_s=.671; p<.05) and RTD (r_s=.930; p<.05), NOR and RTD have also a strong positive relationship (r_s=.696; p<.05).

The findings given in Tables 4, 5, 6, and 7 are summarized and represented, respectively, in Table 8.

Table 8. *Summarized findings for all tasks.*

Tasks	Type of Error	Mean of ED (SD)	Mean of FRD (SD)	Mean of NOR (SD)	Mean of RTD (SD)	Mean of RD (SD)	Mean of GUD (SD)
3	NR	62.08 (27.22)	56.08 (67.24)	2.42 (2.11)	57.33 (65.57)	129 (83.81)	151.33 (73.93)
5	BP	31.42 (24.18)	10.42 (3.23)	6.33 (2.96)	112.83 (66.86)	100.42 (71.19)	129.50 (63.35)
7	CT	16.08 (39.64)	13.33 (8.19)	3.42 (2.61)	36.50 (35.65)	36.33 (14.02)	65.58 (36.28)
9	NF	19 (4.90)	11 (4.43)	1.25 (0.75)	13.75 (25.95)	21.50 (26.14)	33.42 (26.00)
11	RE	4.25 (3.14)	16 (7.46)	4.83 (3.16)	70.33 (57.01)	55 (44.29)	98.58 (52.92)
Mean (SD)		26.57 (19.75)	21.37 (17.47)	3.65 (1.78)	58.15 (33.4)	68.45 (40.26)	95.68 (42.5)

When we inspect the FRD indicator, it can be seen that all participants act similarly except for task 3. In task 3, because of simulating the NR error, we waited for the participants for a while. This is the reason for high FRD for task 3. The mean of the FRD value was determined as 21.37 seconds (SD: 17.47) when all of the FRD values were considered.

The mean of NOR is 3.65 (SD: 1.78) for all errors. This value shows that the participants usually tend to retry 2-5 times when facing a faulty interaction. In addition to this, RTD is another important variable which shows the duration of the retry process. The mean of RTD was determined as 58.15 seconds (SD: 33.4). It means that the participants retried for about 30-90 seconds when they interacted with a fault. On the other hand, the mean of RD, which indicates the mean elapsed time of error detection, was determined as 68.45 seconds (SD: 40.26). Additionally, it can be seen that the participants gave up in 95.68 seconds, on average (SD: 42.5).

For the first two research questions of this study, we searched for an answer by conducting some hypothesis tests. While conducting these tests, we excluded the ED variable of TES, because data collected over this variable would probably change on every distinct system. Besides this, it can be easily stated that data collected over the other variables represent the actions that users perform similarly in all systems after encountering an error. In this way, without the ED variable, the other five variables were considered a tolerance indicator for TES. Regarding this explanation, our null and alternative hypothesis are as below for the first question of this research:

> H_0: There is no difference between the tolerance levels of the participants in the two environments.
>
> H_1: There are some differences between the tolerance levels of the participants in the two environments.

As our observation count (n=7 for E1 and n=5 for E2) is not enough for a parametric test, we used the Mann-Whitney U test in order to compare two different environments' data. Every comparison is based on both the type of error and the data collected over variables individually. The findings are given in Table 9.

Table 9. *P-values of the comparison of E1 and E2 participants.*

THE TYPE OF ERROR	VARIABLES				
	FRD	**NOR**	**RTD**	**RD**	**GUD**
NR	.935	.561	.368	.808	.416
BP	.869	.281	.935	.223	.808
CT	.073	.526	.935	.570	.685
NF	.084	1	.935	.744	.290
RE	.254	.248	.808	.935	.935

p<=.05 is chosen as the significance level for all comparisons.

According to the results in Table 9, H_0 hypothesis cannot be rejected. Although the FRD values of the CT and the NF errors may almost reject H_0, the other twenty-three comparisons show that there is no difference between the tolerance levels of the participants in the two environments.

Because we did not detect any difference between the two environments, we created another null and alternative hypothesis for the second question of the study by considering E1 and E2 together (AE):

H_0: There is no difference between the errors in terms of tolerance indicators (variables).

H_1: There are some differences between the errors in terms of tolerance indicators (variables).

For this analysis, in which we used Friedman test, the indicator values collected from all participants were tested based on all errors, individually. In other words, this test was done in order to detect if there was any difference between the errors in terms of indicator values. The findings from Friedman test are given in Table 10.

Table 10. *P-values of the comparisons for all errors in terms of indicator values.*

INDICATORS				
FRD	**NOR**	**RTD**	**RD**	**GUD**
.05	.000	.001	.001	.000

p<=.05 is chosen as the significance level for all comparisons.

According to the test results in Table 10, it can be seen that H_0 can be rejected for each indicator. From this point, our next step was to find out how the error types differentiate user behavior in terms of indicators, which correspond to the answer to our second research question. In order to determine behavioral differences of the participants, we used Wilcoxon Signed Rank Test. For this analysis, we made ten comparisons between the errors for each indicator. The findings based on the FRD indicator are given in Table 11.

Table 1. *The comparisons based on the FRD indicator.*

Comparisons	Errors	Median	Mean	Z	r	p
1	BP	10.50	10.42	-2.119	0.612	.034
	NR	36.50	56.08			
2	CT	12	13.33	-2.080	0.600	.038
	NR	36.50	56.08			
3	NF	11.50	11	-2.268	0.655	.023
	NR	36.50	56.08			
4	RE	13.50	16	-2.484	0.717	.013
	BP	10.50	10.42			
5	RE	13.50	16	-1.970	0.569	.049
	CT	12	13.33			

p<=.05 is chosen as the significance level for all comparisons.

According to Table 11, except for the RE error, all comparisons made between the NR and the other errors show that the NR error is the error to which the participants had the slowest reaction. In addition to this, the BP and the CT errors caused a faster reaction than the RE error. The analysis results based on the NOR indicator are given in Table 12.

Table 2. *The comparisons based on the NOR indicator.*

Comparisons	Errors	Median	Mean	Z	r	p
1	BP	7	6.83	-3.089	0.892	.002
	NR	2	2.42			
2	RE	4	4.83	-2.200	0.635	.028
	NR	2	2.42			
3	CT	2	3.42	-2.242	0.647	.025
	BP	7	6.83			
4	NF	1	1.25	-2.938	0.848	.003
	BP	7	6.83			
5	NF	1	1.25	-2.352	0.679	.019
	CT	2	3.42			
6	RE	4	4.83	-2.288	0.660	.022
	CT	2	3.42			
7	RE	4	4.83	-3.084	0.890	.002
	NF	1	1.25			

p<=.05 is chosen as the significance level for all comparisons.

Although there is no significant difference between the BP and the RE errors, these two types of errors occupied the participants more than any other error type in terms of recovering from the error. It was also determined that the CT error obligated users to try again more often than the NF error (Table 12 – Comparison 5). As mentioned above, the NOR and the RTD indicators did have a strong positive relationship according to Spearman's ρ test results. Because of the

fact that the findings from the comparisons based on the RTD indicator have similar clues as NORs', we did not consider it necessary to include the findings here. Instead of that, the analysis results based on the RD indicator are given in Table 13.

Table 33. *The comparisons based on the RD indicator.*

Comparisons	Errors	Median	Mean	Z	r	p
1	CT	37.50	36.33	-2.667	0.770	.008
	NR	111.50	129			
2	NF	13.50	21.50	-2.824	0.815	.005
	NR	111.50	129			
3	RE	39.50	55	-1.962	0.566	.050
	NR	111.50	129			
4	CT	37.50	36.33	-1.963	0.567	.050
	BP	109	100.42			
5	NF	13.50	21.50	-3.059	0.883	.002
	BP	109	100.42			
6	NF	13.50	21.50	-2.118	0.611	.034
	CT	37.50	36.33			
7	RE	39.50	55	-2.713	0.783	.007
	NF	13.50	21.50			

p<=.05 is chosen as the significance level for all comparisons.

According to Table 13, although there is no significant difference between the BP and the NR errors, the participants were able to detect these two types of errors later than other types of errors. Additionally, the error type with the fastest detection was found to be NF. In the correlation tests performed between the RD and the GUD parameters, a strong positive relationship was found in all errors except the RE error (r_s=.538; p=.071) for the significance level of .05. For this reason, we did not consider it necessary to include the findings of the GUD indicator, either.

3.2 Qualitative Findings

Qualitative findings were collected from both interviews after the test and from verbal expressions during the test. In the interviews, the participants were first asked what they thought about the mobile website that was being tested. After that, we wanted to obtain their opinions about the effects of the same faults on any website. All collected information, which also corresponds to the answer to our third research question, is divided into three sub-sections (reliability, alternatives, and quality) and presented below:

Reliability

All of the participants stated that the website was insecure. They said that if it was a real-life experience, they would give up shopping at the website. The participant numbered as 5, who thought that the failure was because of his own mobile device, indicated that it might be tried on another device and then he/she would give up if the same errors persisted. The participant numbered as 12 said that "if I knew the website before, I would try to warn website administration. But if it was my first encounter with the website, I would never use it again".

When asked about their thoughts about websites serving other purposes than shopping, they stated that even if it was still insecure, they could be more forgiving. As an example, participants 5 and 10 stated that if it was a website serving a different purpose, the same failures could be ignored.

Alternatives

The participants stated that there were so many alternative websites with a similar purpose. For this reason, they implied that they would prefer an alternative e-commerce website if facing an insecure situation. The participant numbered as 6 said that "Even if it were not a fraud, I would prefer to shop at another e-commerce website". Nevertheless, some views show that the price of the products on the website might affect user behavior. Some comments such as "I can give it one more chance next time but just once", "I can give it one more chance, but I would buy cheap things" and "If the prices are really lower than any other website, I could give a few more chances" can be counted as an example. In the light of these explanations, it can easily be said that prices play an important role in users' tolerance of website faults.

Quality

Participants emphasized that the faults on the website were related to the quality of the website and lack of usability. "More attention should be paid to monetary transactions", "The website seems like an amateur site" and "It does not look nice in terms of professionalism" are some of the comments which need to be considered regarding any kind of website. In addition, some comments questioning the trustworthiness of the website were made as well.

4 Discussion and Conclusion

In this study, we conducted a test in order to detect the differences in users' behaviors and perceptions and investigate users' tolerance of encountering a faulty web interaction while using a manipulated mobile e-commerce website. Instead of detecting the weaknesses of an application with a usability test, it was aimed to enable users to interact with the weaknesses. A shopping scenario which had 11 tasks, including five different faulty tasks, was created for the test. In this way, the changes in users' behaviors, tolerance and perception when encountering a faulty interaction were inspected using Tolerance Evaluation Scale (TES) which had been created for this research by the authors.

From the findings collected from the descriptive analyses, it can be seen that the participants' behaviors are different for each type of error. The interaction duration is longer for the NR error than the others because the system made the participants wait for a while for a response. The FRD value of E2 participants is higher than E1s' (Table 3), but E2 participants behaved more impatiently with regard to avoiding the errors when GUD values are considered (except the NR error). This finding is an important result that shows that internet users are more intolerant in their natural environment. Additionally, Spearman's ρ test results in the CT error also revealed that the participants acted impatiently while they were made to wait for any reason. We believe that this behavior is not limited to a particular error but can also be generalized for any situation that results in having to wait. On the other hand, it was discovered that the participants were also surprised when they understood that they were redirected to the main page unexpectedly instead of reaching the cart page (the RE error).

Similar to the result of the study made by Nah (2004), in this study, it was seen that feedback had a positive effect on the participants' tolerance. When we compare the BP and the NF errors, we can see that whilst the BP error contains a blank page, NF contains an informative "Not Found" feedback. Even if their FRD is similar, there is a big difference between the NOR and

the RTD values. For this reason, it can be said that the feedback message plays an important role in improving user experience. Additionally, we believe that feedback should be given for any kind of faulty situation, but it might also be given for the processes that make users wait.

For the first two questions of the study, we also performed significance tests. One of the important findings is that there is no difference between the tolerance levels of the participants in the two environments, which is the answer to our first question. For the second research question of the study, the important points are summarized below:

- The BP and RE errors are the errors that the participants struggled at most.
- The BP and NR errors are the errors that took the participants the longest to understand as an error.
- The NR error is the error to which the participants showed the slowest reaction (except for the comparison with RE).
- NF is the most quickly recognized error thanks to the effect of feedback.

In the quantitative results, the measured values are considered high due to the participants' psychology of being tested. In other words, it is predicted that users might have lower tolerance if they encounter errors similar to this study in daily life. Even so, in our opinion, the behavioral differences between the error types will be similar in real-life experiences. For this reason, error types and their different effects can be considered as valid.

Mahajan et al. (2016), stated that the service quality and the trustworthiness of a website can be negatively affected by the presence of failures. In the qualitative findings of the present study, similar conclusions were reached. The participants pointed out the importance of the quality of the e-commerce website. On the other hand, they especially emphasized that it does not matter what purpose a website serves; the errors give an amateur impression by damaging usability and professionalism. Even if some of the participants declared that reasonable prices might result in giving an extra chance(s) to an e-commerce website, it is clear that the likelihood of this action is very low.

As a result of this study, it can be stated that creating positive experiences for users depends on knowing how users behave when encountering any type of errors. Generally, on the basis of the findings of the study, our suggestions are as follows:

- Faulty situations can be automatically directed to a page that gives feedback by any system.
- In case of a long process, feedback can be given at certain intervals during the time in which a transaction is performed.
- The types of errors examined in this study might be difficult to track one by one in heavily operating systems. In order to facilitate this process, web mining can be used to detect related/similar error types.
- Another qualitatively obtained information in this study was the price evaluations of the participants. It can be suggested that a newly opened e-commerce site might sell cheaper than other markets in order not to lose its users in case of possible errors in the recognition process.

Thanks to this study, we observed how users behave when encountering some kind of faulty web interaction. Even though we were not able to use eye-tracking glasses in this study for technical reasons, using these kinds of devices will probably present important clues about user

behaviors. The presented findings of the study might be helpful for system designers and academics for future work.

ORCID

Veli Özcan Budak http://orcid.org/0000-0002-0960-0542

Emre Akadal http://orcid.org/0000-0001-6817-0127

Sevinç Gülseçen http://orcid.org/0000-0001-8537-7111

References

Avizienis, A. (1978). Fault-tolerance: The survival attribute of digital systems. *Proceedings of the IEEE*, 66(10), 1109–1125. https://doi.org/10.1109/PROC.1978.11107

Begosso, L. C., & Filgueiras, L. V. L. (2006). Human error simulation as an aid to HCI design for critical systems. In *Proceedings of VII Brazilian Symposium on Human Factors in Computing Systems*, (pp. 120–127). ACM. https://doi.org/10.1145/1298023.1298040

Card, D. N. (1998). Learning from our mistakes with defect causal analysis. *IEEE Software*, 15(1), 56–63. https://doi.org/10.1109/52.646883

Everard, A., & Galletta, D. F. (2005). How Presentation Flaws Affect Perceived Site Quality, Trust, and Intention to Purchase from an Online Store. *Journal of Management Information Systems*, 22(3), 56–95. https://doi.org/10.2753/MIS0742-1222220303

Galletta, D. F., Henry, R., McCoy, S., & Polak, P. (2004). Web Site Delays: How Tolerant are Users?. *Journal of the Association for Information Systems*, 5(1), Article 1. https://doi.org/10.17705/1jais.00044

Graham, I. (2003). *A Pattern Language for Web Usability.* Pearson Education.

Guse, D., Schuck, S., Hohlfeld, O., Raake, A., & Möller, S. (2015). Subjective quality of webpage loading: The impact of delayed and missing elements on quality ratings and task completion time. In *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*, (pp. 1–6). IEEE. https://doi.org/10.1109/QoMEX.2015.7148094

Heckel, R., & Mariani, L. (2005). Automatic Conformance Testing of Web Services. In M. Cerioli (Ed.), *Fundamental Approaches to Software Engineering* (pp. 34–48). Springer. https://doi.org/10.1007/978-3-540-31984-9_4

ISO. (1998). *Ergonomics of human-system interaction*. (9241-11). Geneva: International Organization for Standardization.

Laprie, J.-C. (1995). Dependable Computing and Fault Tolerance: Concepts and Terminology. In *Twenty-Fifth International Symposium on Fault-Tolerant Computing, 1995, ' Highlights from Twenty-Five Years'*, (pp. 2–11). IEEE. https://doi.org/10.1109/FTCSH.1995.532603

Lazar, J., Meiselwitz, G., & Norcio, A. (2004). A taxonomy of novice user perception of error on the Web. *Universal Access in the Information Society*, 3(3), 202–208. https://doi.org/10.1007/s10209-004-0095-9

Ma, L., & Tian, J. (2007). Web error classification and analysis for reliability improvement. *Journal of Systems and Software*, 80(6), 795–804. https://doi.org/10.1016/j.jss.2006.10.017

Mahajan, S., Li, B., Behnamghader, P., & Halfond, W. G. J. (2016). Using Visual Symptoms for Debugging Presentation Failures in Web Applications. In *2016 IEEE International Conference on Software Testing, Verification and Validation (ICST),* (pp. 191–201). IEEE. https://doi.org/10.1109/ICST.2016.35

Meyers, S. (2004). The most important design guideline? [User interfaces]. *IEEE Software*, 21(4), 14–16. https://doi.org/10.1109/MS.2004.29

Munson, J. C., Nikora, A. P., & Sherif, J. S. (2006). Software faults: A quantifiable definition. *Advances in Engineering Software*, 37(5), 327–333. https://doi.org/10.1016/j.advengsoft.2005.07.003

Nah, F. F.-H. (2004). A study on tolerable waiting time: How long are Web users willing to wait? *Behaviour & Information Technology*, 23(3), 153–163. https://doi.org/10.1080/01449290410001669914

Nielsen, J. (1993). *Usability Engineering*. Morgan Kaufmann.

Norman, D. (2013). *The Design of Everyday Things: Revised and Expanded Edition*. Basic Books.

Ocariza, F., Bajaj, K., Pattabiraman, K., & Mesbah, A. (2013). An Empirical Study of Client-Side JavaScript Bugs. In 2013 ACM / IEEE International Symposium on Empirical Software Engineering and Measurement, (pp. 55–64). IEEE. https://doi.org/10.1109/ESEM.2013.18

Phalgune, A., Kissinger, C., Burnett, M., Cook, C., Beckwith, L., & Ruthruff, J. R. (2005). Garbage in, garbage out? An empirical look at oracle mistakes by end-user programmers. In *2005 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC'05),* (pp. 45–52). IEEE. https://doi.org/10.1109/VLHCC.2005.40

Ploski, J., Rohr, M., Schwenkenberg, P., & Hasselbring, W. (2007). Research issues in software fault categorization. *ACM SIGSOFT Software Engineering Notes*, 32(6), 6–es. https://doi.org/10.1145/1317471.1317478

Ramsay, J., Barbesi, A., & Preece, J. (1998). A psychological investigation of long retrieval times on the World Wide Web. *Interacting with Computers*, 10(1), 77–86. https://doi.org/10.1016/S0953-5438(97)00019-2

Rubin, J., Chisnell, D., & Spool, J. (2008). *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests*. Wiley.

Shneiderman, B., & Plaisant, C. (2004). *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Addison Wesley.

Tan, Z., Zhu, J., Chen, J., & Li, F. (2019). The Effects of Response Time on User Perception in Smartphone Interaction. In T. Z. Ahram & C. Falcão (Eds.), *Advances in Usability, User Experience and Assistive Technology* (pp. 342–353). Springer. https://doi.org/10.1007/978-3-319-94947-5_34

Tzeng, J.-Y. (2004). Toward a more civilized design: Studying the effects of computers that apologize. *International Journal of Human-Computer Studies*, 61(3), 319–345. https://doi.org/10.1016/j.ijhcs.2004.01.002

Copyright © 2020 by the author(s). Licensee Prague University of Economics and Business, Czech Republic. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution License (CC BY), which permits use, distribution and reproduction in any medium, provided the original publication is properly cited, see http://creativecommons.org/licenses/by/4.0/. No use, distribution or reproduction is permitted which does not comply with these terms.

The article has been peer-reviewed.

Editorial record: First submission received on 23 July 2020. Revisions received on 24 August 2020. Accepted for publication on 31 August 2020. The editor in charge coordinating the peer-review of this manuscript and approving it for publication was Zdenek Smutny.