ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF SCIENCE ENGINEERING AND TECHNOLOGY

PROTOTYPE FIDELITY AND USER EXPERTISE IN USABILITY TESTING: A STUDY WITH PORTABLE NAVIGATION DEVICE

M.Sc. THESIS

Gamze KAYA KAPLAN

Department of Industrial Design

Industrial Design Programme

MAY 2015

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF SCIENCE ENGINEERING AND TECHNOLOGY

PROTOTYPE FIDELITY AND USER EXPERTISE IN USABILITY TESTING: A STUDY WITH PORTABLE NAVIGATION DEVICE

M.Sc. THESIS

Gamze KAYA KAPLAN
(502101906)

Department of Industrial Design

Industrial Design Programme

Thesis Advisor: Assoc.Prof.Dr. Şebnem TİMUR ÖĞÜT

MAY 2015

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

KULLANILABİLİRLİK TESTLERİNDE PROTOTİP UYGUNLUĞU VE KULLANICI UZMANLIĞI: TAŞINABİLİR NAVİGASYON CİHAZI İLE BİR ÇALIŞMA

YÜKSEK LİSANS TEZİ

Gamze KAYA KAPLAN
(502101906)

Endüstri Ürünleri Tasarımı Anabilim Dalı

Endüstri Ürünleri Tasarımı Programı

Tez Danışmanı: Doç. Dr. Şebnem TİMUR ÖĞÜT

MAYIS 2015

To my family,

FOREWORD

First of all, I would like to point out my appreciation to my advisor Assoc.Prof.Dr. Şebnem TİMUR ÖĞÜT for her support and encouragement. I would also like to state my deepest gratitude to my co-advisor Asst.Prof.Dr. Erdem DEMİR for his patient guidance and support during the development of this study and encouragement even when I got lost in the depth of the subject.

My thanks also go to the twenty participants for their time and motivation to take part in the usability tests. I would also express my special thanks to Esin Arsan for her contribution to my analysis and Saniye Fışgın for her support during the process.

I want to thank my parents, who supported me in every way during my studies and special thanks for their encouragement

Finally, I would like to express my deepest gratitude and love to my husband, Ilke Kaplan, who has been my best friend for half of my life and walks me through this journey offering endless love, patience, motivation and support.

May 2015 Gamze Kaya Kaplan

TABLE OF CONTENTS

Page

ABBREVIATIONS

ISO **:** International Organization for Standardization
PND **:** Portable Navigation Device
POI **:** Point of interest

LIST OF TABLES

Page

LIST OF FIGURES

Page

PROTOTYPE FIDELITY AND USER EXPERTISE IN USABILITY TESTING: A STUDY WITH PORTABLE NAVIGATION DEVICE

SUMMARY

Usability of a product, interface or service is a crucial issue in terms of success of design. Several evaluation methods were introduced in the usability literature. Usability testing is the most reliable assessment method compared to other methods such as usability inquiry and usability inspection. Usability tests are must-have parts of the design process. They are more practical to cover the real-world experience and can be applicable to any phase of the design process of physical (consumer products; from cars to chair, computers to lamps…etc.) and digital products (such as; websites, apps, any business software…etc.). Even if the main purpose and outcomes of the usability testing differ according the studied object, in general, the main aim is to identify usability problems.

Several factors can change the data provided by usability tests. Based on the *Four-Factor Framework of Contextual Fidelity Model* (Sauer et al, 2010), these factors are system prototype, user characteristics, test environment and task scenarios. In recent years, studies have been done to find the effects of these factors individually, especially with prototyping fidelity and user expertise. Beside these, quite a few studies in the literature were found for the effect of the combination of these factors. Only one research with a physical product was conducted for both prototype fidelity and user expertise effects on usability testing outputs (Sauer et.al 2010). Although the results of the related research provide some insight about the effects of these two factors, it is not adequate to predict the possible effects especially on digital interfaces.

The purpose of this thesis is to contribute to the literature by looking at the influence of prototype fidelity and user expertise on usability testing outputs of a digital interface and interaction between these factors if any. The main contribution with these revelations will be providing knowhow for those who want to design specific usability tests. To simulate the realistic experience and let participants to complete the task without any interruption, "Retrospective Think Aloud Protocol" and "Performance Measurements" were used to gather data to achieve the desired goals. Twenty participants were used in total divided into four groups with five participants each; 5 novice and 5 expert participants worked with the low fidelity prototype (paper) while another 5 novice and 5 expert participants worked with the high fidelity prototype (the device itself). By doing this evaluation, usability problems were analyzed in terms of number, severity and variety of types. In addition, performance data (time on task, success rate) is also evaluated. Accordingly, it becomes possible to comment on which user groups takes part actively in what kind of prototypes.

Although the sample size of five participants per each group has little statistical power, the results showed that, novice users found significantly more usability problems than experts in total and the difference between novices and experts was higher under low-fidelity prototype rather than high fidelity one. The study also showed that in order to understand the differences between the tested groups (novices and experts) and interfaces (low and high fidelity), usability problems must be isolated from each other. By doing this, it is important to consider on the cause of the problem to define types and subtypes for the classification with descriptions in detail. According to the result of the study, significant differences were found between user groups with the categories; use flow, menu categorization, interactive components and aesthetic & visual. In addition, content related problems were found significantly more in low fidelity prototypes and the interaction between expertise and fidelity is found significant with in the category of "aesthetic and visual". Lastly, experts showed better performance with the time spent on test.

KULLANILABİLİRLİK TESTLERİNDE PROTOTİP UYGUNLUĞU VE KULLANICI UZMANLIĞI: TAŞINABİLİR NAVİGASYON CİHAZI İLE BİR ÇALIŞMA

ÖZET

Bir ürünün, arayüzün ya da servisin kullanılabilirliği tasarımın başarısı açısından önemli bir konudur. Bu bağlamda kullanılabilirlik literatüründe çeşitli değerlendirme metotları sunulmuştur. Kullanılabilirlik testleri tasarım süreçlerinin vazgeçilmez bir parçası olup diğer yöntemlerle karşılaştırıldığında en güvenilir değerlendirme metodu olarak karşımıza çıkmaktadır. Çünkü kullanıcı testleri yüksek oranda yüzeysel gerçekliğe sahiptir. Yani, her türlü fiziki (tüketici ürünleri; sandalyeden arabaya, lambadan bilgisayara… vs. kadar) ve dijital (web sitelerinden aplikasyonlara ve şirket yazılımlarına… vs. kadar) ürünün tasarımının herhangi bir aşamasında gerçek deneyimleri otaya koyan pratik bir yöntemdir. Kullanılabilirlik testlerinin çıktıları her ne kadar üzerinde çalışılan ürüne ve çalışmanın içeriğine göre değişse de asıl amaç kullanılabilirlik problemlerini ortaya çıkarmaktır.

Kullanılabilirlik testlerinin çıktıları üzerinde etki eden birçok faktör bulunmaktadır. Sauer ve diğ. (2010) tarafından tasarlanan Dört Faktör Bağlamsal Uygunluk Modeline (the Four Factor Framework of Contextual Fidelity Model) göre, bu faktörler; sistem prototipi, kullanıcı özellikleri, test ortamı ve test senaryoları olarak gruplanmıştır. Son yıllarda, bazı araştırmacılar bütün bu faktörlerin etkilerinin ayrı ayrı incelendiği çalışmalar yapmışlardır. Prototip uygunluğu ve kullanıcı uzmanlığının kullanılabilirlik testlerine birlikte etkisi sadece fiziki bir ürün (yer temizleme makinesi) ile yapılan bir çalışmada incelenmiştir (Sauer ve diğ., 2010). İlgili çalışmanın sonuçları bu iki faktörün etkileri hakkında bazı anlayışlar sağlasa da, çalışmanın sonuçları bu iki faktörün özellikle dijital ara yüzler üzerindeki olası etkilerini tahmin etmek için yeterli değildir.

Bu çalışmanın temel amacı prototip uygunluğunun ve kullanıcı uzmanlığının kullanıcı testleri çıktıları üzerindeki etkilerini ve birbirleri arasındaki ilişkiyi bir dijital ürün olan taşınabilir araç navigasyon cihazı üzerinden inceleyerek literatüre bu konuda katkı sağlamaktır. Bu çıktılar sayesinde kullanıcı testi hazırlayanlar için temel bir bilgi kılavuzu oluşturulması hedeflenmiştir.

Kullanılabilirlik testleri genel olarak tasarım süreçlerinde, kullanıcıların geribildirimlerinden beslenerek ürünlerin problemli kısımlarının ortaya çıkarılması, değerlendirilmesi ve geliştirilmesinde aktif rol oynamaktadır. Bunun yanı sıra daha çok tasarım sürecinin tamamlanmasından sonra yapılan kullanılabilirlik değerlendirmelerinde karşımıza çıkan ürün performans değerlendirmesi de kullanılabilirlik testleri aracılığı ile yapılabilmektedir.

Kullanılabilirlik problemlerinin tespiti için çalışmanın içeriğine ve amacına bağlı olarak çeşitli yöntemler kullanılmaktadır. Bu çalışmada, kullanıcılardan ürün

kullanımıyla ilgili bilgi toplamak için hem uzman gözlemleri hem de kullanıcı yorumları değerlendirmeye alınmıştır. Özellikle kullanıcı yorumları için geçmişe yönelik yüksek sesli düşünme (retrospective think aloud) yöntemi kullanılmıştır. Bu yöntemde kullanıcılar ile her işlem sonrasında işlem sırasında karşılaşılan problemler detaylı olarak değerlendirilmiştir. Bunun yanı sıra performans ölçümü yöntemi de kullanılmıştır.

Bu çalışmada kullanıcılara taşınabilir navigasyon cihazı ile daha çok sürüş öncesi süreçte yaptıkları 6 işlem verilmiştir. Bu işlemler, kayıtlı adrese rota hazırlama, ev adresini güncelleme; favori adres ekleme; mevcut rotaya ara adres ekleme; kısa yol menüsü oluşturma; harita ekranı bilgi okuma-güncelleme; hız limiti sesli uyarı ayarı gibi görevleri içermektedir. Toplamda 20 kişi her birinde 5 kişinin bulunduğu 4 gruba ayrılmıştır; 5 deneyimsiz ve 5 deneyimli kullanıcı düşük uygunluklu prototip (kağıt) ile; diğer 5 deneyimsiz ve 5 deneyimli kullanıcı ise yüksek uygunluklu prototip (ürünün kendisi) ile çalışmıştır. Her kullanıcı grubu aynı işlemleri gerçekleştirmiş olup yapılan testler ortalama otuz dakika sürmüştür ve daha sonra değerlendirilmek üzere kaydedilmiştir.

Kullanılabilirlik problemleri, kullanıcı testlerinin en önemli çıktılarıdır. Bu çalışmada, bulunan kullanılabilirlik problemleri miktar, önem dereceleri ve tip çeşitliliklerine göre analiz edilmiştir. Problemlerin önem dereceleri, düşük, orta ve yüksek olmak üzere üç seviyede değerlendirilmiştir. Problem tip çeşitlilikleri için ise öncelikle bir kategori listesi hazırlanmıştır. Hazırlanan bu liste kullanılabilirlik testleri sonrasındaki problemlerin içerikleri temel alınarak tekrar revize edilmiştir ve problemler ilgili kategorilere ayrılmıştır. Bu problem kategorileri; "içerik, sayfa yapısı, kullanım akışı, menü kategorizasyonu, interaktif elemanlar, sistem durumu-geribildirimler ve estetik-görseldir. Ek olarak, kullanıcı ve prototip grupları arasındaki performans verileri (başarı oranı ve işlemler için harcanan süre) değerlendirilmiştir. Bu analizlerden sonra hangi kullanıcı grubunun hangi prototip tipinde daha çok data sağlayarak aktif rol aldığını belirtmek mümkün olabilmiştir.

Literatürde, incelenen örneklem sayısının, her grup için 5 kişi, istatistiki olarak anlamlı sonuçlar elde etmek için yetersiz olduğu belirtilmektedir. Ancak bu çalışmanın sonucunda, kullanıcı uzmanlığı ve prototip uygunluğunun kullanılabilirlik testi çıktılarına etkilerini gözlemleyebildiğimiz istatistiki olarak anlamlı sonuçlar da elde edilmiştir.

Çalışmanın en önemli bulgularından birisi, deneyimsiz kullanıcıların deneyimlilere göre toplamda istatistiki olarak önemli ölçüde daha çok kullanılabilirlik problemi ortaya çıkarmış olmasıdır. Bu iki grup arasındaki fark düşük uygunluklu prototipte daha fazladır.

Kullanıcı grupları ve prototip tipleri arasındaki farkları daha iyi analiz edebilmek için, ortaya çıkarılan problemler sebepleri de dikkate alınarak yapılan gruplamaya göre ayrı ayrı değerlendirilmiştir. Bu gruplamanın sonuçları göstermektedir ki; ortalamalara bakıldığında en fazla içerik ile ilgili problem ortaya çıkarılmıştır. Özellikle düşük uygunluklu prototiple çalışan kullanıcılar içerik ile ilgili önemli ölçüde daha fazla problem ortaya çıkarırken, ortalama sayılar göz önünde bulundurulduğunda kullanım akışı ve sayfa yapısı ile ilgili problemler, bu prototipte daha fazla bulunmuştur. Yüksek uygunluklu prototip ile çalışanlar ise estetik- görsel kategorisinde önemli ölçüde daha fazla problem ortaya çıkarmıştır. Ek olarak, menü kategorizasyonu, interaktif elemanlar, sistem durumu-geribildirim kategorilerinde de

ortalama problem sayıları göz önünde bulundurulduğunda yüksek uygunluklu prototipte daha fazla problem bulunmuştur.

Kullanıcı uzmanlığı değerlendirmeye alındığında, deneyimsiz kullanıcıların deneyimlilere göre menü kategorizasyonu, interaktif elemanlar ve estetik-görsel kategorilerinde önemli ölçüde daha fazla problem buldukları gözlemlenmiştir. Bunun yanı sıra, deneyimli kullanıcılar, kullanım akışları kategorisinde deneyimsizlere göre önemli ölçüde daha fazla problem bulmuştur.

Performans değerlendirme sonuçları ise deneyimli kullanıcıların deneyimsizlere göre işlemleri yaparken harcadıkları süre özelinde daha iyi performans sergilediğini göstermiştir. Bu sonuca ek olarak, ortalama değerler göz önünde bulundurulduğunda; deneyimli kullanıcıların genel olarak işlemleri tamamlarken daha başarılı oldukları gözlenirken, kullanıcı uzmanlığı ve prototip uygunluğunun kullanıcıların işlemleri tamamlama süreleri üzerinde daha az etkisi olduğu ortaya çıkmıştır. Son olarak, prototip uygunluğu ve kullanıcı uzmanlığı arasındaki etkileşim ise önemli ölçüde estetik- görsel kategorisindeki problemlerde gözlemlenmiştir.

Bütün bu sonuçlar göstermektedir ki; temel amaç kullanılabilirlik problemlerini bulmak ise düşük uygunluklu prototipler yüksek uygunluklular kadar ve deneyimsiz kullanıcılar ise deneyimliler kadar efektiftir. Eğer temel amaç bir arayüzün yapısını değerlendirmek ise; çalışmanın kapsamı ve içeriği de göz önünde bulundurulduğunda kullanıcıların uzmanlık seviyeleri ve prototip uygunluk derecesi gözetilmeksizin kullanıcı ve prototip seçimi yapılabilir. Ek olarak performans ölçümünün temel amaç olarak belirlendiği çalışmalarda ise deneyimli kullanıcılar ve yüksek uygunluklu prototiplerle çalışmak daha uygundur.

1. INTRODUCTION

Usability of a product, an interface or a service is a crucial issue in terms of quality and success of design and interaction between the product and the user has been discussed for several years. If users have no difficulties to use a system and satisfy with the process, it makes the system successful (URL-1). Otherwise, they may tend to give up using it.

As it is proposed by Preece et al. (2002), interaction design process can be divided into four main parts:“ (1) identifying needs and establishing requirements, (2) developing alternative designs, (3) building interactive versions of the designs and (4) evaluating designs” (p 169). Usability is an important issue in all these parts. In each part, different methods are used to address usability issues. Real users usually come to play important roles in the fourth part, i.e. evaluation of the replicas built in the third part.

There are a lot of usability evaluation methods in the literature. Partala and Kangaskorte (2009) mentioned in their article that the current usability evaluation methods can be categorized as usability testing, usability inspection and usability inquiry. The main objective of these methods is to obtain usability problems. Among them, usability testing is the most reliable method in terms of higher face validity, that is to say, it gives results that is closer to the real life experiences. This method has been used “almost two decades in interaction design field” (Partala and Kangaskorte, 2009). Usability testing gives an opportunity to evaluate what will happen when the product reaches the real users (Dumas and Redish, 1999). In other words, prior to launching a product, it provides researchers direct information about the way of using the system. Moreover, studying with users under realistic experiences instead of imaging usage scenarios helps one to find unpredictable problems that cannot be discovered during evaluation.

Although usability testing is the most referred and popular method, it also comes with some constraints. A test setting does represent participants, products, scenarios

and test environments. The quality of test outputs directly depends on the test setting. Sauer et al. (2010) identified four factors (presented in "*the Four-Factor Framework of Contextual Fidelity"* and explained in detail in section 2.4) within this formulation that can influence the outcomes of the usability testing: system prototype, testing environment, user characteristics and task scenarios. All these factors can influence user behavior and satisfaction during the test (Sauer et al., 2010). For example, if the participants does not represent the target group, the results can be irrelevant; if the prototype is not understood by participants, users may have confusion to address the problems; if the level of noise is higher in the test environment, it may negatively affect the performance of users; if the task scenarios are not formulated regarding to real ones (depth and breadth), the test does not provide relevant results.

The quality of the test is the conformance of the outputs (usually the depth & breadth of the usability problems identified during the test) to the goals of the test. In this study, we are focusing on the formative usability evaluation. That is to say, usability tests are done to identify usability problems and improve the general usability of a designed product. In this case, it is important to look into the number of the usability problems, the severity of the usability problems that are identified and the variety of problem types.

In this thesis, we will focus on the prototype fidelity and user expertise, because these two factors essentially interact with each other and directly influence the output quality. The main aim is to provide knowhow for these factors to prepare guidelines to contribute the design process of usability tests. It is difficult to specify arguments and prepare such guidelines for other two factors; test environment (field or lab) and task scenarios (breadth and depth of a task scenario). There are some rules to prepare tasks scenarios, but they only help to create general structure and these scenarios mostly depend on the context of the case. Similarly, determining a test environment is also case specific and hard to define general guidelines.

There a lot of studies focusing on each factor (expertise and fidelity) on the test output quality independently. There is only one study on a physical product with quite simple interface (Sauer et.al 2010). Although the results of the related research provide some insights about the effects of these two factors, it is not adequate to predict the possible effects especially on digital interfaces.

1.1 Purpose of Thesis

With this study, it is expected to contribute to the literature by looking at the influence of prototype fidelity and user expertise on usability testing outputs of a digital interface and interaction between these factors if any. The main contribution with these revelations will be providing knowhow for those who want to design specific usability tests.

Do we really need fully interactive and visually perfect prototypes (which are required more effort) to understand the system is usable or not? Do experts always perform well and provide all data we needed or is there any uncovered data that we can also get from novices? In order to answer these questions we need to be specific about the success of the usability test. In this context, it is focused on the usability tests, which are done to improve the usability in the product development process. The success of the usability test mostly related with the output quality of the test. In this case the number of the usability problems, the severity of the usability problems that are identified and the variety of problem types are defined as the quality indicators.

Taking these questions above as a basis, the following sub questions were formulated:

In what way does the fidelity of prototypes influence the participants' ability to reveal usability problems in usability testing?

- Do users reveal more problems on high fidelity prototypes?
- Are revealed problems more severe on high fidelity prototypes?
- Do users show differences to find the variety of problem types on both low and high fidelity prototypes?

In what way does the participants' expertise influence the revealed usability problems in usability testing?

- Do expert users reveal more usability problems on both low and high fidelity prototypes?
- Are revealed problems by expert users more severe on both low and high fidelity prototypes?

- Do both novice and expert users show differences to reveal the variety of problem types?

Is there any interaction between fidelity and expertise throughout the usability testing outputs?

Beside these main questions, additional questions were formulated to compare performance data between user groups and prototypes. In formative studies, objective usability metrics; such as task completeness and time on task may also be measured, but these are not in the main focus also in this study. These questions are;

In what way does the fidelity of prototype influence the performance data?

- Do users always spend less time on high fidelity prototypes?
- Are users more successful to complete the task on high fidelity prototypes?

In what way does the participants' expertise influence the performance data?

- Do experts always spend less time to complete the task?
- Are experts more successful to complete the tasks?

1.2 Structure of Thesis

As it was already said above, the main aim of this research is to discover the influence of prototype fidelity and user expertise on usability testing outputs. In order to conduct this research, this thesis was formulated with five main chapters. Figure 1.1 shows the logical structure of this thesis.

In chapter 1, the topic, purpose and structure of the thesis and the research questions are presented. In chapter 2, the literature is reviewed consisting of four main parts. First part starts with the usability evaluation and brief descriptions of evaluation methods in usability literature. Second part, investigates the selected method "usability testing" and third part introduces the outputs of usability testing. In last part, the test setting factors that influence the usability testing outputs and the main variables of this study are covered; user expertise and prototype fidelity were discussed in detail referring to the previous research. In Chapter 3, the design and issues regarding the conduct of the usability test is presented. In Chapter 4, the results and analysis of the study are pointed out. Finally, in chapter 5, the conclusions

for the whole research study are drawn with answering the research questions to be responded throughout this study followed by recommendations for future studies.

Chapter 1	**Introduction**				
Chapter 2	**Literature Review**				
	Usability Evaluation	Usability Testing	Outputs of usability testing	Test Setting Factors that influence outputs	
Chapter 3	**Design and Conduct of the Research**				
	Usability test	Test Materials	Participants	Test environment	Procedure
Chapter 4	**Results and analysis**				
	Analysis of performance data	Analysis of usability problems	Discussion		
Chapter 5	**Conclusions and Recommendations**				

Figure 1.1 : Diagram of structure of this thesis

2. LITERATURE REVIEW

In this chapter, current literature about usability is elaborated to construct the conceptual background of this thesis. In the first section, usability and usability evaluation methods are overviewed. In the second section, usability testing as an evaluation tool in usability is investigated in detail. In the third section, the outputs of usability testing and output quality are described. Lastly, in the fourth section, the test setting factors that influence the usability testing outputs are described, overviewing the research literature about two main factors, prototype fidelity and expertise in usability testing.

2.1 Usability Evaluation

Usability, a main term in product and user interaction has been discussed for decades. Shackel(1981) provides the first fully discussed and detailed formal usability definition as "[a system's] capability in human functional terms to be used easily and effectively by the specified range of users, given specified training and support, to fulfill a specified range of tasks, within the specified range of environmental scenarios" (as cited in Bruno and Al-Qaimari ,2004, p.1).

In usability literature, several definitions of usability have been proposed by different researchers. The most widely referred definition of “usability” is given in ISO 9241-11. ISO defines usability by adding the user perception as: "The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use." (ISO, 1998). In this definition, *effectiveness* is defined as “the accuracy and completeness with which users achieve specific goals”. The definition is made for *efficiency* as the “resources expanded in relation to the accuracy and completeness with which users achieve specified goals” and for *satisfaction* as the “freedom from discomfort and positive attitudes towards the use of the product.” (ISO,1998). These metrics are explained in the section 2.3 in detail.

The classification of Nielsen (2012a) is also widely accepted among usability experts. He defined usability as a "quality attribute" of an interface and helps to discover whether the usage is easy or hard. Nielsen also pointed out that a usable product is easy to learn, remember and has low error rate. With these properties, he offered five quality components; learnability, efficiency, memorability, errors and satisfaction (Nielsen, 1993).

2.1.1 Usability evaluation methods

Usability evaluation methods are frequently used in different stages of the product development process. Umar and Tatari (2008) identified three stages of the product development as" before, during and after" and discussed the usability evaluation methods in a relation to these stages. Usability evaluation methods were categorized by Scriven (1967) into two main groups in terms of their objectives; summative and formative (as cited in Sonderegger, 2010). In the development stages of "before" and "after", summative methods are used. Summative methods aim to determine "overall quality of a finished product" by testing the "performance requirements" and compare the alternative designs (Umar and Tatari, 2008). Beside this, formative methods are used "during" the development stage and the main objective is to improve the usability of the design by identifying the usability problems (Umar and Tatari, 2008). A diagram was proposed by Umar and Tatari (2008) to define the relation between this evaluation methods and stages (before, during and after) of product development (Figure 2.1).

In summative studies, normally inquiry methods are used to collect quantitative data and these studies focus on quality and efficiency of the final product and comparison of alternative designs before the design process (Hartson, Andre and Williges, 2001). The most popular inquiry methods are; contextual inquiry, user questionnaires, user interviews, focus groups and field observations and logging actual use.

In formative studies, both usability inspection and usability testing methods are used. These methods throughout the whole development phase to obtain user feedbacks and focuses on usability problems uncovered during the design process. "Usability inspection is the generic names for a set of methods that are all based on having evaluators inspect a user interface."(Nielsen, 1995b). Usability inspection methods are divided into two main groups; (1) methods that performed by single evaluator;

heuristic evaluation, heuristic estimation, cognitive walkthrough, feature inspection, standards inspection and (2) methods that performed by group of evaluators; pluralistic walkthrough and consistency inspection. In addition, usability tests also conducted for summative purposes to gather performance related data, such as; task completion success and time on task. The main difference of summative test is "to evaluate a product through defined measures, rather than diagnosis and correction of specific design problems, as in formative evaluation." (URL-2)

This thesis focused on the second stage of the product development. In this case, the main thing is to evaluate the design process to improve the usability. Thus, formative methods are in the focus of this study.

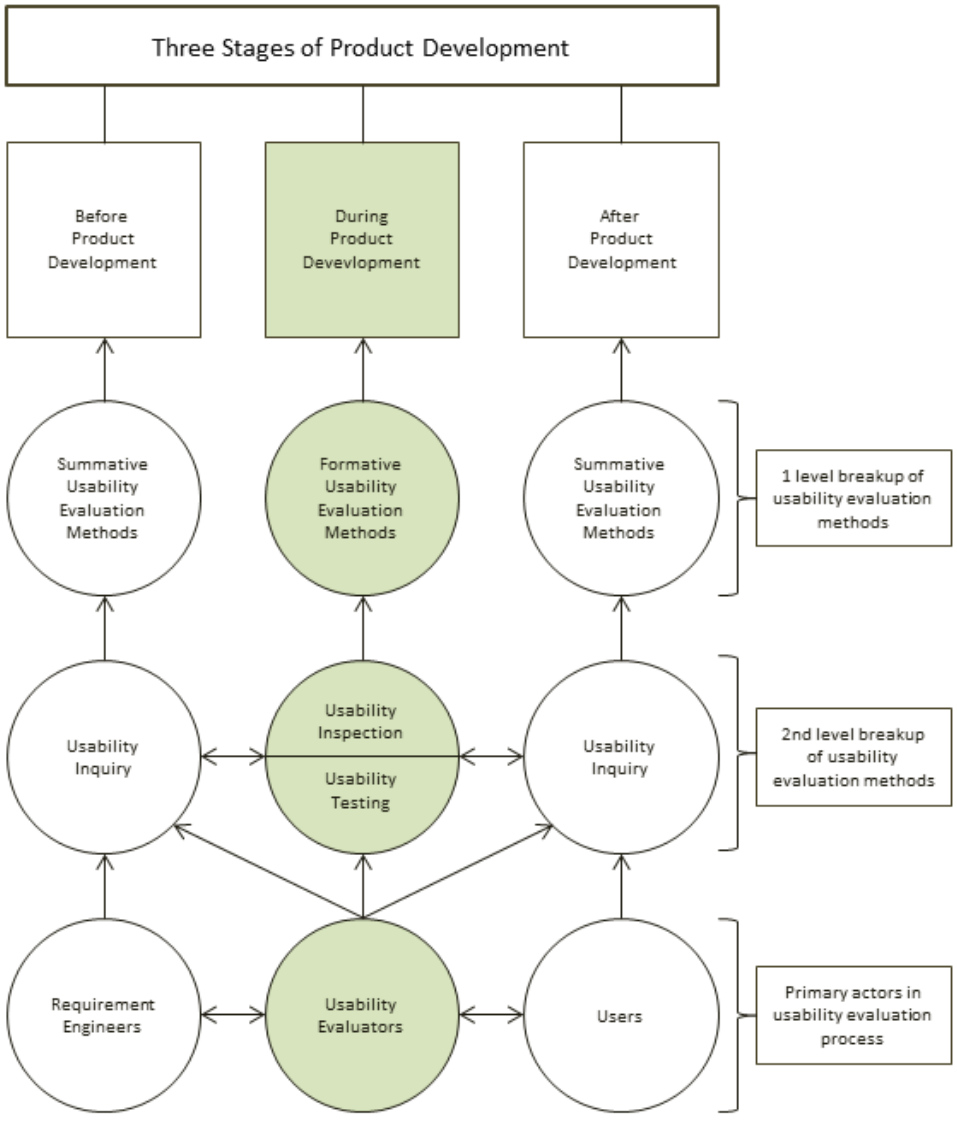

Figure 2.1 : Diagram of "Conceptual Visualization of Usability Evaluation Process" (Umar and Tatari, 2008).

Among the methods used for formative usability evaluation, usability testing is superior to usability inspection methods, because they are most reliable method in terms of higher face validity, that is to say, it gives results that is closer to the real life experiences. "Usability testing employs techniques to collect empirical data of the interaction of representative end users with the product by completing representative tasks in a controlled environment." (Sonderegger, 2010, p.5). It provides researchers direct information about the way of using the system by realistic experiences and helps to find some unpredictable problems that they cannot discover during the evaluation. Because of this superiority, the focus of this thesis will be on usability testing.

All these evaluation methods focus on the collection of usability problems while detecting the user's demands, experiences and priorities (Sonderegger, 2010). Each method has some advantages and disadvantages. Table 2.1 shows a comparison table proposed by Genise (2002).

2.2 Usability Testing

Usability testing is widely accepted method and has been used for almost two decades in interaction design field (Partala and Kangaskorte, 2009). This method can be applied to several hardware products (consumer products; from cars to chair, computers to lamps…etc.) and software products (such as; websites, apps, any business software…etc.) at any phase of the design process.

Dumas and Redish (1999) defined the five characteristics of usability testing. These characteristics are; (1) improving the usability of the product (primary goal), (2) representation of the real users as participants, (3) doing real tasks by participants, (4) observing and recording all the reactions and (5) comments of the participants and analyzing all the recordings to detect the usability problems (p.22).

Usability testing gives an opportunity to evaluate what will happen when the product gets to the real users (Dumas and Redish, 1999). During usability testing, while the participants use the test object to complete the desired task, the sessions are recorded to use in later analysis such as; usability problems, performance data (task completion, time on tasks …etc.) and perceived usability. To detect the usability problems, different ways of converting a usability tests are offered depends on the context, procedure and aim of the study. These methods are; Thinking aloud

Table 2.1 : Comparison table proposed by Genise "Usability Evaluation: Methods ad Techniques: Version 2.0" (2002)

Evaluation Method	Evaluation Method Type	Applicable Stages	Description	Advantages	Disadvantages
Testing	**Think aloud protocol**	Design, coding, testing and release of application	Participants in testing express their thoughts on the application while executing set tasks	Less expensive Results are close to what is experienced by users	The Environment is not natural to the user
	Remote Usability testing	Design, coding, testing and release of application	The experimenter does not directly observe the users while they use the application though activity may be recorded for subsequent viewing	Efficiency, effectiveness and satisfaction, the three usability issues, are covered	Additional Software is necessary to observe the participants from a distance
Inquiry	**Focus groups**	Testing and release of application	A moderator guides a discussion with a group of users of the application	If done before prototypes are developed, can save money Produces a lot of useful ideas from the users themselves Can improve customer relations	The environment is not natural to the user and may provide inaccurate results. The data collected tends to have low validity due to the unstructured nature of the discussion
	Interviews	Design, coding, testing and release of application	The users are interviewed to find out about their experience and expectations	Good at obtaining detailed information Few participants are needed Can improve customer relations	Cannot be conducted remotely Does not address the usability issue of efficiency
Inspection	**Cognitive walkthrough**	Design, coding, testing and release of application	A team of evaluators walk through the application discussing usability issues through the use of a paper prototype or a working prototype	Good at refining requirements Does not require a fully functional prototype	Does not address user satisfaction or efficiency The designer may not behave as the average user when using the application
	Pluralistic walkthrough	Design	A team of users, usability engineers and product developers review the usability of the paper prototype of the application	Usability issues are resolved faster Greater number of usability problems can be found at one time	Does not address the usability issue of efficiency

Protocol (users verbally express thoughts during test), Question-asking Protocol (moderator ask some questions to the user), Shadowing method (an usability expert explains to moderator the actions of users), Coaching Method (during the test, user is free to ask questions), Teaching Method (novice user learn from expert user), Co-discovery Learning (user couple works together), Performance Measurement (the test session is recorded with a software or another camera), Log File Analysis (a moderator analysis the usage data), Retrospective Testing (a type of think-aloud protocol that users and moderators reviews the records together). Even if the way that they collect the data is different, all of them are aiming to get usability problems to fix them and increasing the usability of the system (Ivory, 2001).

As mentioned before, usability tests are most reliable method in terms of higher face validity, that is to say, it gives results that is closer to the real life experiences prior to launching a product. It also provides researchers direct information about the way of using the system by realistic experience and helps to find some unpredictable problems while studying with users that they cannot discover during the evaluation. However, in order to fully utilize these advantages, because of some reasons, inaccurate results can be reported. Table 2.2 presents the values and limitations of usability testing which was included in the study produced by group of students from Miami University (2004)

In spite of the limitations and reasons for resistance, usability testing is the most effective evaluation method to uncover usability problems if the study is conducted well.

Table 2.2 : Values and limitations of usability testing (Students of Miami University, 2004)

Values for companies	Minimize cost. Minimize risk. Companies acquire a competitive edge. Increase revenue, product sales, and brand loyalty. Create a historical record of usability benchmarks for future
Values for product developers and designers	Provide more efficient time Minimize the need for unscheduled updates. Make developing documentation and training easier.

Table 2.2 (continued) : Values and limitations of usability testing (Students of Miami University, 2004)

Values for users	Focus on developing usable products Increase user satisfaction
Limitations and resistance	Not always represent the real environment Does not necessarily prove that products work May include test participants who do not represent the target user May be costly Not always the best technique to use Extends the product development lifecycle

2.3 Outputs of Usability Testing

It is expected to contribute to the literature that the influence of prototype fidelity and user expertise on usability testing outputs of a digital interface and interaction between these factors if any. The main contribution with these revolutions will be providing knowhow for those who want to design specific usability tests.

In line with the main objective of this thesis, it is crucial first to be specific about the quality of the usability test outputs. The success of a usability test should be defined in relation to the specific goal of the usability test. Thus, to better analyze the usability of the interface, we first need to identify the type of usability evaluation that we are planning to conduct. Sauro (2011) identified four types of usability evaluation that can be read as the main goals;

- *Detecting usability problems in an interface:* To focus on to find problems that users have and to find and fix the problematic area in the interface. Most usability studies are based on this purpose.
- *Estimating a parameter:* To focus on to identify parameters of the interface such as; completion rate of all users, the average task time, and the perception of usability.
- *Making a comparison:* To compare two or more interfaces to find which has higher completion rates, shorter task times or higher satisfaction scores.
- *Comparing to a Benchmark:* To compare the parameters such as completion rate provided by the real use of interface with the Benchmark scores.

The main focus of this thesis depends on the first type; *Detecting usability problems in an interface* (number, severity and the variety of problem types). As it is mentioned in the evaluation categorization by Sauro (2011), the main purpose of most usability tests is to find what usability problems that users come up with the interface. Usability test are also conducted for summative goals; *Estimating a parameter* (success rate, time on task).

In the product development process, most usability studies cover the combination of these evaluations. For summative goals, there are three main metrics to measure usability. These measures are *effectiveness, efficiency* and *satisfaction*. *Effectiveness* is defined as "the accuracy and completeness with which users achieve specific goals"; *efficiency* as the "resources expanded in relation to the accuracy and completeness with which users achieve specified goals" and *satisfaction* as the "freedom from discomfort and positive attitudes towards the use of the product." (ISO, 1998). With the same approach as ISO, Nielsen (2001) offered the most basic measures based on the definition of usability as a quality metric: success rate (whether users can perform the task at all), the time a task requires, the error rate and users' subjective satisfaction. Beside these measures, Nielsen also offers that evaluating the navigation path of the users may provide some useful data for the discovering navigational problems (2001).

Sonderegger (2010) defined the measures in usability test in his study in five categories. These measures are; performance data, perceived usability, user emotions and user experience, physiological measure and usability problems. Satisfaction, also defined as perceived usability, measures how users satisfied with the product or interface. Satisfaction is usually collected with questionnaires, semi structured interviews and evaluating user behaviors during the test. Evaluating user emotions requires some special evaluation tools and more expertise on emotional researches. In addition, there are also some specific measures needed to record physiological data with heart rate data (HR), galvanic skins response (GSR) and blood volume pressure (BVP) to realize the users' reactions (Sonderegger, 2010).

In this thesis, measures for user emotions, the specific physiological measures and measure for perceived usability are not included based on the purpose. Because the main focus of this study is to find usability problems and to analyze these problems into defined categories and severity scales. Beside this, the performance data of the

participants will be also measured to get quantitative insights about user's behaviors according to interface.

2.3.1 Usability problems

In formative usability evaluations, the main output of the usability evaluation is a set of usability problems. With identifying these problems, designers and developers could fix the problematic area and develop the interface.

Usability problems are defined in ISO 9241-11 as "problems that influence the effective, efficient, and satisfactory use of the system in a specified context of use" (ISO, 1998). Lavery et al. described the usability problem in their article as an "aspect of the system and/or demand on the user which makes it unpleasant, inefficient, onerous or impossible for the user to achieve their goals in typical usage situations" (1997, p.254). In usability literature, there are several definitions of usability problems in the context of the related study. Usability problems are generally defined as mistakes by users in completing a task. These mistakes occur in some conditions such as when a user does not understand or misunderstand of functions, elements, utterances or actions of the system by users (Mäuselein, 2007).

It is important to begin with the analysis the cause of the problems and the effects on users. These effects are defined in the website of in Human Computer Interaction at Virginia Tech as; (1) psychological effects (e.g. confusion, irritation), physical effects (e.g. hand-eye coordination, fatigue), (2) perceptual effects (e.g. inability to discern text, inability to notice an object) and (3) task- related effects (e.g. inability to complete a given task) (URL-3).

Tullis and Albert (2008) with the same approach of the effects mentioned above defined some examples of usability issues in their book showed in Figure 2.2.

- Anything that prevents task completion
- Anything that takes someone off course
- Anything that creates some level of confusion
- Anything that produces an error
- Not seeing something that should be noticed
- Assuming something should be correct when it is not
- Assuming a task is complete when it is not
- Performing the wrong action
- Misinterpreting some piece of content
- Not understanding the navigation

Figure 2.2 : Examples of usability issues by Tullis and Albert (2008)

In this thesis, all these approaches guided to recognize the problems with the studied design. In a usability test, usability problems can be identified through two sources (1) observations of experts or (2) participants verbal reports. To observe the usability problems, it is needed to focus on the participants' verbal expressions such as; confusion, frustration, dissatisfaction, pleasure or surprise and nonverbal behaviors such as facial expressions and/or eye movement (Tullis and Albert, 2008).

While evaluating usability problems, it is also important to catch when the problem begins-ends and if the problem causes other problems in the current task session. Some problems may cause a task failure, some may cause another problem, and some are identified by users and just cause confusion for a while (Sonderegger, 2010).

When the main goal is to improve usability of a system, it is crucial to detect usability problems. The success of the usability test mostly related with the output quality of the test. In this case, the success is defined as with the number of the usability problems, the severity of the usability problems that are identified and the variety of problem types appear as the main indicators for the quality of the test outputs. It is important to set the usability test considering on what the practitioners are trying to discover. Based on the purpose of this thesis, it is crucial to see if there is a difference in variety of problems identified by user groups under prototypes with different fidelity.

2.3.1.1 Number of usability problems

The input of the problems is reported by the expert's observations and the participant's verbal expressions. The data "Number of problems" gives the relation between the amounts of problems found by each participant and total reported problems.

Some problems can be reported only by one participant while some of them can be found by all of them. To present the number of usability problems, beside the total number of reported problems, it is also important to define how many distinct problems were found in the interface.

2.3.1.2 Severity of usability problems

In usability studies, in addition to report all problems one by one, it is more valuable to identify range of the problems according to their severity.

According to Nielsen (1995a), severity is defined as a combination of three factors; frequency, impact and persistence. These three factors of severity were described below:

- *Impact*: how much trouble will affect user's experience?
- *Persistence*: how many times will a user experience the problem?
- *Frequency*: how many users will be affected by the problem?

Hertzum (2006) also mentioned in his article that, the evaluators are expected to define the impact and persistence of the usability problems beside the descriptions of them.

In literature, some researchers are still discussing the relation between severity and frequency. Sauro (2014) analyzed nine different studies which all measured the correlation between frequency and severity; some of them found strong relation, others found no relation. The severity of the problem based on the judgments of the evaluator and the participants (not always be grouped homogenously) and it affects the problem identification. Thus the frequency and severity are not always correlated.

Although the frequency and persistence of problems can be assessed objectively by direct measurement, the assessing of the problem requires a subjective approach (Sauro, 2014). The most common method for assessing severity of a usability problem is rating. For last few decades, different severity rating systems have been offered. Even if the scales and wordings are different, the structures of the approaches are similar. The main finding will be whether the problem has minor or major effect on users (Hertzum, 2006).

Nielsen (1995a) determined the severity with 4 point scale. He firstly describes "0" as the problem that is not classified. Rating of "1" represents cosmetic problem and it can be fixed if there is enough time. Rating of "2" represents minor usability problem and has low priority. Rating of "3" is major usability problem and has high priority, thus it is important to fix. Lastly, rating of "4" is defined as "usability catastrophe" and the problems in this category must be fixed before product can be released.

The following scale by Rubin (1994) also including 4 scales for problem severity. The most important scale is “4” and represents that the affected part of the design is not able to use by users because of the problematic design. The scale “3” represent the problem is severe and the problem limits the usage of the product. Moderate problem rated as “2” and user should make moderate effort to overcome the problem. The scale “1” named irritant and the problem occurs not so often and also defined as cosmetic problem (as cited in Sauro, 2013)

Another study conducted by Dumas & Redish (1999) and they also offered more task related 4-point scale for severity ratings but they started with the rating low as Level 4 to high as Level 1. If the problem prevents completion of a task, the problem is rated as Level 1. Some problems can create significant delay and frustration and these problems are classified as Level 2. The third scale named as Level 3 and this group of problems has minor effect on usability. The last scale is Level 4 and problems in this category have minor effect and can be fixed in the future (Dumas and Redish, 1999).

The last reviewed severity categorization offered by Sauro (2013). They used 3 severity scales instead of four because of the difficulty to distinguish easily between levels 2 and 3. The first level is defined as minor and the problem is rated with this level if it causes some hesitation or slight irritation. The second level categorized as moderate and includes the problems that make some users fail the task, causes some delays and moderate irritation. The third level is critical and this group of problems cause extreme irritation and lead to task failure. In addition, they also have another category to collect the data as insight, suggestion and positive comment.

In this thesis, with the same approach as Sauro (2013), the three-point scale was used to categorize the severity of reported problems. The offered three-point scale with the following categories presented in Table 2.3

For this analyze, it is better to use more than one evaluator to categorize and define the severity ratings of the usability problems for more reliable and valuable results. Nielsen (1995a) stated in his article that, to increase the quality mean of the severity ratings, more evaluators are needed and for many practical purposes rating results from three evaluators is satisfactory. Beside this, Macnamara (2005) also mentioned in his article that to achieve maximum reliability, two or more coders should be used.

In this study, two usability experts (one of them is the author of this thesis) were used to rate severity because of the time and resource limits.

Table 2.3 : Usability problem severity scale

Low severe	Minimal usability problem, doesn't prevent user to complete task, can be fixed easily
Medium severe	Moderate usability problem, affects on user to complete task, takes time to be fixed
High severe	Major usability problem, mostly cause task failure, important to be fixed unless take more time before the product released

2.3.1.3 Variety of usability problem types

Usability problems must be diagnosed and described properly and isolated from each other for qualified results. In line with the main goal of a usability tests, it is important to understand the causes of identified problems in order to make a step further in solving them.

Nielsen and his colleague Molich were developed a set of heuristics to evaluate the interface usability in 1990 (as cited in Nielsen, 1995c). Afterwards, to be better understood by usability researchers and experts, with the analysis of 249 problems based on the factor analysis, Nielsen refined these heuristics (1995c). Nielsen's ten heuristics is presented in Figure 2.3

Nielsen's ten usability heuristics		
	Visibility of system status	Recognition rather than recall
	Match between system and the real world	Flexibility and efficiency of use
		Aesthetic and minimalist design
	User control and freedom	Help users recognize, diagnose and recover from errors
	Consistency and standards	
	Error prevention	Help and documentation

Figure 2.3 : Ten usability heuristics by Nielsen (1995c)

These heuristics mostly refers specifications that an interface should have and does not give the causes of the problems. There are some specific categories offered by researchers using these heuristics of Nielsen (1995c). Partala and Kangaskorte (2009) defined six categories related to the discovered usability problems. These are; (1) communication using metaphors (e.g. meaning of the icon could not be inferred), (2) choice of concepts (e.g. the information could not be associated with the label), (3) interaction styles (e.g. the scrollable field could not be noticed), (4) media interface

(e.g. the file started to play automatically), (5) navigational structure of the application (e.g. complexity of the path), and (6) navigation and information presentation (e.g. hyperlinks could not be noticed). This categorization is too narrow to be applied in other studies.

Another categorization offered by Mäuselein (2007) divided in seven; consistency, distinct graphics, locality, system response, system structuring, user orientation and wording. These categories do not cover all the aspects of the product.

Travis (2014) published a usability guideline for website evaluation and it includes nine main categories and related guidelines, which is important to evaluate the usability of the interface. This guideline gives several tips to evaluate; home page usability, task orientation, navigation and IA, forms and data entry, trust and credibility, writing and content quality, page layout and visual design, search usability, help, feedback and error tolerance.

In this study, the main focus is to find the problems related with product aspects. With the analysis of the previous studies, and taking the general guidelines into account, firstly the problems were defined in six main types trying to refer the aspects of the product. This first categorization will be used to define causes of the problems at the beginning of analysis. These six types and related sub-types are presented in Table 2.4. After usability tests, based on the discovered problems, the types, subtypes and definitions of the problems were reformulated. The final categorization is presented in section 4.1.3.1.

Table 2.4 : Variety of usability problem types (First categorization)

Type	Subtype
Content (visual &textual)	Complexity in content
	Unclear information
	Unclear wording&abbreviations
	Unclear iconography
	Inconstance information
	Inconstancy in information quality

Table 2.4 (continued) : Variety of usability problem types (First categorization)

Type	Subtype
Page Layout	Improper positioning of components regarding the task steps
	Improper functional grouping in the page
	Inconsistent page layout
	Sequence of interactive components
Information Architecture	Unclear menu categorization
	Improper subcategories
	Duplicated menu items
	Inappropriate number of task steps
	Unconvenient connections between related pages
	Improper depth / length
Interactive components	Unclear input and input format
	Unclear interactive components
	Unclear interaction in lists/tables
	Lack of interactive components
	Incorrect usage of interactive components
	Inadequate auto complete
System status and response	Inadequate & unnecessary feedback
Aesthetic and visual	Inappropriate color usage
	Inappropriate text usage
	Inappropriate image usage
	Visual complexity
	Improper visual hierarcy

2.3.2 Performance data

In usability evaluation, the objective ratings; effectiveness and efficiency are defined as performance data for product usability.

2.3.2.1 Effectiveness

Effectiveness is the first objective metric and measured by the completion rate of given tasks by user (Sonderegger, 2010). The completion of given task is called as "success rate" and it is an easy way to understand and document the usability of the interface by measuring the users' ability to complete tasks (Nielsen 2001). Nielsen also mentioned in his study that it is more effective if we talk with numbers while presenting the usability of any product (2001). Thus, percentage of the task completion is a basic way to explain the success rate of users.

Success rate is basically defined as if the user completes the task correctly or fails. In addition, there is also a sub-situation and it is named as "partial success" in most of the studies. According to Nielsen (2001), the users who complete much of the task should not have the "zero" score as the users who did nothing and failed. In his article, he also mentioned that the severity of the user error effects to score partial success.

In this thesis success rate was divided into three categories; *success* (if the participant completed the task directly without an help), *success with help* (if the participant complete the task with indirect guidance of moderator) and *failure* (if the participant complete the task with the direct guidance of moderator)

2.3.2.2 Efficiency

The second objective metric of the performance data is efficiency. To measure efficiency, it is needed to look deeper into user behaviors. The typical measures of efficiency proposed by Jordan (1998) collecting some data points from user behaviors; deviation from the critical path (e.g., number of unnecessary clicks during task completion), error rates (e.g., number of clicks on the home or back button before task completion), and time on task (e.g., time needed to accomplish the task) (as cited in Sonderegger, 2010)

In this study, "time on task" metric was used for efficiency measure. The data observed as deviation from critical path and error rates were analyzed in detail as usability problems.

2.4 Test Setting Factors That Influence Outputs

Usability practitioners should be aware of that the usability testing is a simulation and is not perfectly represents the real usage situation (Sonderegger, 2010). The quality of the representative situation determines the quality of the output.

There are four principle components of the human-machine system framework offered by researchers Bennett (1972, 1979) and Eason (1981): user, task, tool and environment (as cited in Sonderegger, 2010). Sauer et al. (2010) offered an improved framework based on these four factors as a guideline to conduct usability testing (as cited in Sonderegger, 2010). Figure 2.4 shows this framework, named *the Four-Factor Framework of Contextual Fidelity.* This framework identifies the factors (system prototype, testing environment, user characteristics and task scenarios) that influence the outcomes of the usability testing (Sauer et al., 2010). As their explanation; they derived this framework from three main sources: "(a) previous models that addressed the issue of fidelity in usability testing (explained in the article in detail), (b) pertinent issues discussed in the usability literature (e.g. user competencies), and (c) issues that play a role in ergonomics beyond the usability literature (e.g. physical and social environment) (Sauer et al., 2010, p.130).

This framework contains four main factors and related subordinate factors which refer to different aspects of fidelity for each factor. Sometimes the fidelity of these four factors may not represent the final usage situation and these fidelity differences influence the user behavior and satisfaction during the test while threating the reliability and validity of the usability test (Sauer et al., 2010).

The first factor; *testing environment* includes physical features (e.g. the location; laboratory of field, the size of the laboratory, noise levels…), social features (e.g. other humans such as observers) and application domain. For instance, if the level of noise is higher in the test environment, it may negatively affect the performance of users.

The second factor; *task scenarios* divided into the sub-factors; breadth (the complexity degree of the modelled task environment) and depth (the level of detail to complete a particular task). As an example, if the task scenarios are not formulated regarding to real ones (depth and breadth), the test does not provide relevant results.

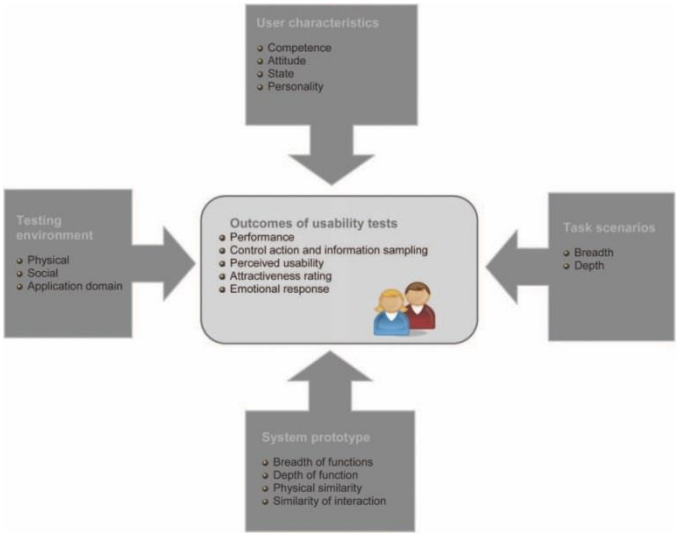

Figure 2.4 : Four-Factor Framework of Contextual Fidelity (Sauer et al., 2010)

For the *prototype factor*, the authors offered four sub-factors; breadth of functions (scope of features), depth of functions (fidelity of functions), physical similarity and similarity of interaction. For example; if the prototype is not understood by participants, users may have confusion to address the problems.

The last factor is *user characteristics* also divided into four sub-factors; competence (knowledge, skills and abilities), attitude (e.g. environmental concern, openness towards technology), state (e.g. mood) and personality (anxiety or extraversion) (Sauer et al. 2010). To better understand the effects of this factor, it is needed to look into the characteristics of target groups. In this case, if the participants do not represent the target group, the results can be irrelevant.

All these factors can guide the usability practitioners to find what influence to the outputs of the usability testing. In usability literature, there have been several studies conducted to evaluate these factors empirically regarding the importance and context of those studies. Sonderegger (2010) overviewed that, in usability testing, studies mostly have been conducted over the factors; fidelity of prototypes, test environments and user expertise which play central role in the product development. There are a lot of studies focusing on each factor on the test output quality independently. Many studies conducted just to evaluate one individual factor such as user expertise (Dillon and Song, 1997; Ziefle, 2002; Faulkner and Wick, 2005; Gerardo, 2007; Shluzas et al., 2013) and testing tool (prototype) (Tam, 2006; Mäuselein, 2007; Sauer and Sonderegger, 2009; Lim et al., 2006; Magnussen, 2010; Virzi et al, 1996; Walker et al, 2002; Sefelin et al., 2003). But, there is only one

study on a physical product with a quite simple interface (Sauer et.al 2010). Although the results of the related research provide some insights about the effects of these two factors, it is not adequate to predict the possible effects especially for on digital interfaces.

In this thesis, we will focus on the prototype fidelity and user expertise influence on a digital interface. These two factors essentially interact with each other and directly influence the output quality. The main aim is to provide knowhow for these factors to prepare guidelines to contribute the design process of usability tests. Because, it is difficult to specify arguments and prepare such guidelines for other two factors, test environment (field or lab) and task scenarios (breadth and depth of a task scenario). There are some rules to prepare task scenarios, but it only helps to create general structure and these scenarios mostly depend on the context of the case. Similarly, determining a test environment is also case specific and hard to define general guidelines.

2.4.1 Prototype fidelity

Prototyping is an essential feature in the design process. In order to understand the problems of the design, we need prototypes prior to product release.

Beaudouin-Lafon and Mackay (2002) defined the term prototype as "a concrete representation of part or all of interactive system" (p. 1007). Lim et al. (2008) defined prototypes as "representative and manifested forms of design ideas". Prototypes can be implemented quickly (Nielsen, 1993) and provide advantages to save both time and money (Lundberg, 2010). Prototypes are also effective ways to communicate with customers, development teams and users to understand their demands. In the interaction design context, prototypes give designers an opportunity to examine the information architecture, interactive elements, content and basic visual aesthetics of the system before the finishing of the product. In addition, the process of prototyping also provides feedbacks to correct possible problems in the system. Prototypes are design decisions and have various representation levels from easy to make sketches such as paper prototypes to highly interactive computer-based prototypes (Mäuselein, 2007).

Mäuselein describes the usage of prototypes in four phases in design process (2007). The first phase is requirement specification, which is most significant process in

design. Users can easily articulate their requirements and give feedbacks about what they need or do not need through prototypes. The second phase is representation in user testing. Prototypes are obligatory objects in user tests and by representing the real system, users can simulate the real tasks to experience the usability of the system. The third phase is application for iterative design. Nielsen (2011) defined iteration in his article that "iteration simply means to step through one design version after another." By conducting usability evaluation on each step, users can study with revised versions based on the usability findings from old ones. Prototypes play a significant role in this iterative design process because they are easy to modify. The last phase is communication and documentation. As it is already mentioned above that, prototypes are an effective communication tool between designers, users, customers and developers.

Prototyping has two dimensions, vertical and horizontal prototyping. Dumas and Redish (1999) mentioned in their research that, in vertical prototypes; only small set of the features are integrated and only few of them have deep functionality to simulate a realistic user test. The researchers also defined horizontal prototypes that included wide range of features with little or almost none functionality.

In usability literature, prototype is one of the key features in the test setting. It has been shown that, there are different ways of conducting usability tests, with a prototype in the spectrum that has less representative prototypes on one side and final products on the other side. Usability tests can be conducted with any prototypes in this spectrum and this spectrum refers "fidelity".

There are different definitions of the term "fidelity". Sauer et al. defined the fidelity as "the degree to which a model of the system resembles the target system refers to the fidelity of the model. The fidelity of the model (or prototype fidelity) may considerably, ranging from a low-fidelity simulation of the system (e.g., paper prototype) to a fully operational prototype, which is (almost) identical to the real system." (2008). Magnussen (2010) defines the fidelity as the level of detail in a prototype. The most common usage for prototype categorization basically defined in two levels; high-fidelity and low-fidelity. Beside this categorization, McCurdy et al. (2006) offer five dimensions to define a prototype beside the definition of low- or high fidelity. These dimensions are; visual refinement (look of the prototype), breadth of functionality (scope of features), depth of functionality (fidelity of

functions), richness of interactivity (similarity of interaction) and richness of data model (actual data).

2.4.1.1 Low fidelity prototypes

Sefelin et al. (2003) describe the low fidelity prototyping as “the visualization of design ideas at very early stages of the design process”. Low fidelity prototypes basically represent the product but the material that is used to produce a prototype is not same as the final product. They are easy to prepare, usually made by paper based material and easily editable. In addition, they are often recommended to use at the beginning of the development process (Lundberg, 2010). They usually called as mockups because of the simple and rough representations of a design (Magnussen, 2010). Even if they are not fully interactive, they are visually similar to the final product and still give an idea about the tested system. On the contrary, because of the low functionality, working with low fidelity prototype sometimes can be disadvantageous. Sauer and Sonderegger (2009) mentioned in their article that users can mentally anticipate of the appearance of the prototype and this drawback can be effective on their ratings.

Low fidelity prototypes are made with physical materials such as, paper, whiteboards, or chalkboards (Petrie and Schneider, 2007). The most common usage of low fidelity prototypes are “Paper prototypes” and it has been used in usability evaluations since the early 1990’s (Tam, 2006). Figure 2.5 presents an example of paper prototype of a mobile application. (URL- 4)

"Paper prototyping is a variation of usability testing where representative users perform realistic tasks by interacting with a paper version of the interface that is manipulated by a person 'playing computer' who doesn't explain how the interface is intended to work” (Snyder, 2003, p.4)

Figure 2.5 : Paper prototype (URL-4)

2.4.1.2 High fidelity prototypes

Preece et.al (2002) described high fidelity prototyping that "uses materials that you would expect to be in the final product and produces a prototype that looks much more like the final thing."(p. 245). High fidelity prototypes are also functional and interactive and with color usage, they are more close to the finished design and provide realistic conclusions when being evaluated.

In this context, high fidelity prototypes are very effective to convey user-interface specifications and system behaviors, thus detailed specifications can be read and understood by development team easily (Virzi et al., 1996). Beside these advantages, some problems can occur with high fidelity prototypes. For instance, it takes too much time to build and a prototype can set some expectations that are hard to change. Sometimes designers and developers refuse to make changes (Rettig, 1994), because more resources such as; time and money are needed to produce a prototype. Rettig (1994) also indicated that, because the high fidelity prototypes are fully functional and interactive, it is important to control all possible bugs before the user test not to block the task.

In interaction design context, high fidelity prototypes are computer-based prototypes that are written with scripting languages and often developed by applicants using interface builders (Petrie and Schneider, 2007).

2.4.1.3 Previous comparative studies on the effects of prototype fidelity

Prototype fidelity has been much researched to find the differences between the effects of different fidelities in usability testing. The researches mostly in interaction design context and consider on low (paper, interactive computer) and high (interactive computer, html based) fidelity prototypes and the differences between evaluated data from usability tests.

Tam (2006) focused on whether the similarities and differences in types of usability issues (occurrence of confusion, experience problems, deviation from the path, page coherence, and screen update impact) and task success. The researcher worked on an online book review community web application; used paper as a low prototype medium and Html coded version as a high fidelity prototype medium. Figure 2.6 presents sample pages of these prototypes. Six users were randomly selected from the population who actively participate in online communities. She used both

quantitative and qualitative (thematic analysis) methods to obtain the results. With quantitative analysis, the results showed that users showed similar success to complete the tasks under both prototypes. According to thematic analysis, she found the number and the types of problems are similar under both prototypes. Users were confused more with the paper prototype and she concluded that, paper prototype was not effective as web prototype according to thematic analysis method.

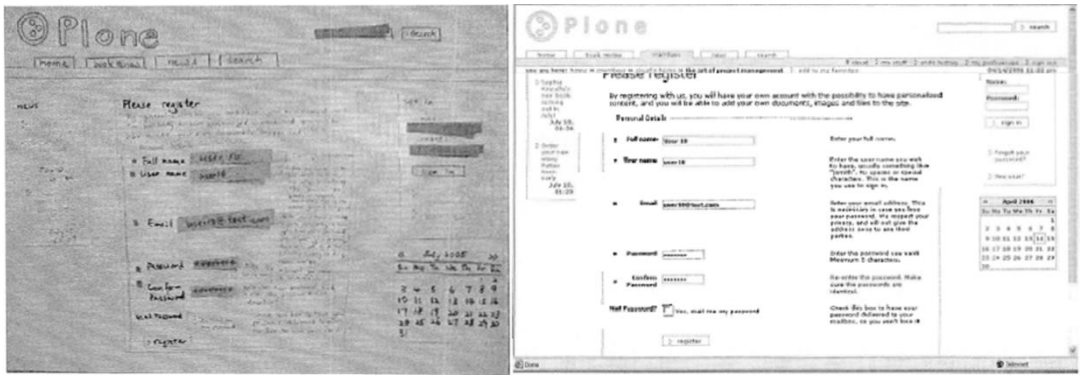

Figure 2.6 : Low and high fidelity prototypes of online book review community web application (Tam, 2006)

Mäuselein, (2007) also compared paper and computer prototypes in a user test with the study on an online media center. The main goal of the study was the comparison of the prototype efficiency. By doing this study, the researcher evaluated feedbacks and suggestions into usability problems (number, severity and types), comments and ideas to find which and how the types of prototypes affect the user participation in usability tests. Beside this, success rate were also analyzed. Ten participants were used and they switched the prototype after completing tasks on the first one. Users made almost twice as many comments and suggestions on paper prototype because they were in free task of describing screen in paper prototyping session. There is no difference with the number and types (position, consistency, distinct graphics, locality, system response, system structuring) of uncovered usability problems, but more severe problems were found with computer based prototype. In addition, users were more successful to complete tasks under computer-based prototype.

Beside to focus only on fidelity, Lim et al. (2006) also tried to find the medium effect on a mobile device. They studied on mobile device and used a finished product as the high-fidelity prototype and build one paper and one computer-based prototype as low-fidelity. They used 15 participants divided 5 for each prototype. The study conducted to analyze usability problems (number and types). The result showed that

users found more problems on high fidelity prototype and less problems on low fidelity paper based prototype. Even they found similar problems in all prototypes; especially there are differences between high fidelity (finished product) and computer-based low-fidelity version. The researchers also pointed out that the representation type of prototype is also important in usability tests.

In a similar way, Walker et al. (2002) conducted a study to understand the influence of the fidelity and medium of the prototype on the uncovered usability issues (severity and type). For this study, they prepared both high (paper and computer) and low fidelity (paper and computer) prototypes of two online banking websites. 28 participants were used and after conducting the tests, comments made by participants are categorized into usability issues (like or dislike for aspects of the website; problems navigating through the site to complete the assigned tasks; suggestions for site improvements; confusion about the site). The usability issues were rated for severity by the researchers and ten outside judges. They found that, there was no significant difference between low and high fidelity prototype regarding total usability issues and severity of usability issues. The types of usability issues found were significantly different between low and high fidelity prototypes. Beside these results, the medium effect were also analyzed and found that participants made more comments about computer prototypes rather than paper ones.

Studies focused not only the quantitative analysis, but also made some analysis from subjective preferences from users. Sefelin et al. (2003) studied whether the users' willingness to criticize the system and suggestions for its improvement are affected by using low fidelity -paper and computer- prototypes. They used two systems (calendar system and touch screen ticket machine) and for each system, they developed computer and paper prototypes with similar functionality. 24 participants were used in total. The results indicated almost same quantity and quality of critical user statements (functions, operational design, behavior and screen layout and wording) from the usability tests of both prototypes. In addition, they also evaluated the medium of prototypes and found users preferred computer based prototypes which they felt more freedom to moving around and exploring the interface and also felt less observed than paper prototype which manipulated by moderator. Also more graphical related comments were provided from computer prototypes.

In usability literature, it has been mentioned in many studies that low fidelity prototypes are often accepted in the early design process. On the contrary, Virzi et al. (1996) addressed another question in their studies: "In the later stages of user-interface design, are low-fidelity prototypes as effective as high-fidelity prototypes in identifying usability problems?" (Virzi et al., 1996) analyzed whether the low fidelity prototypes are effective as high fidelity prototypes at the later stages to detect usability problems. They studied on two different products, a portable electronic-book and interactive voice response system and for each study; they used both low- and high-fidelity prototypes. For the first study, they used the device itself as a high fidelity prototype and simulated screens of the device on a paper as a low fidelity prototype. For the second study, the high-fidelity prototype was built using TLFXTM software and, the low fidelity prototype was simulated by an experimenter reading aloud the possible responses instead of computer based on the subject's input. Twenty participants were used in each study. They found that, substantially similar amount and types of problems were found in both low- and high-fidelity prototypes.

As a summary, low fidelity prototypes are effective as the high fidelity ones to discover usability problems in terms of number and types of the problems. In these reviewed studies, types of problems were defined according to context of the study thus; these results are not been generalized. Beside this, users made more comments on high fidelity prototypes and they found more severe usability problems. In addition, high fidelity prototypes are better than low ones when the performance metrics (success rate, completion time) are measured in usability tests because users have more confusion on paper prototypes used as low fidelity.

2.4.2 User characteristics

Warell (2001) defined the user as "any individual who, for a certain purpose, interacts with the product or any released element (system, part, component, module, feature, etc., manifested in software or as concrete objects) of the product, at any phase of the product life cycle" (as cited in Liu et al., 2010)

In usability literature, there are some definitions for participant groups and categorizing users related to concept of the research (Faulkner and Wick, 2005). The main purpose of this categorization is to better realize and understand the real users' behaviors. Participants are representing the real users' characteristics such as, users'

age, gender, education level, technical needs, cultural background, and attitude and skill level (Liu et al., 2010).

To choose appropriate users for studies is always required more focus to gather valuable and reliable data from usability tests. All these factors have an importance based on the context of the study. Some of them are not always be applicable to all studies because test users should be good replicas of future users of the product.

Sauer et al. (2010) focused on the user characteristics as influencing factor on usability outputs. They offered four category for user characteristics; competence, attitude, state and personality. The researchers mentioned that the *competence* is the most important user characteristic that based on knowledge, skills and abilities of users (Sauer et al, 2010).

According to these definitions, the term *competence* refers to the term *user expertise* that is used in this thesis. The main theme of this thesis is to use participants with different expertise levels in usability test to investigate the differences on test results. Thus, the literature on *user expertise* is overviewed in the following section.

2.4.2.1 User expertise

In the context of usability testing, participants are generally categorized as (1) novice or (2) expert. The definition of novice and expert also depends on the purpose of the research. According to Shneiderman's (1992) perspective, there are three kinds of users: "(1) novice users – users who know the task but have little or no knowledge of the system, (2) knowledgeable intermittent users - users who know the task but because of infrequent use may have difficulty remembering the syntactic knowledge of how to carry out their goals, (3) expert frequent users – users who have deep knowledge of tasks and related goals, and the actions required to accomplish the goals." (as cited in Liu et al. 2010).

Nielsen (1993) offered three main concepts for users' expertise; (1) general knowledge and familiarity with computers; (2) understanding of the task domain and (3) expertise in using the system. Computer usage experience in general represents length of usage and familiarity with computer and the purpose of usage. The domain of user is also important if the tested system is used for specific purpose such as; software for users with a background in computer programming. The last approach is

user's experience with system and gives information about how long and how much a person has used a system with two levels as novice and expert usage.

Shneiderman's (1992) definition of users is based on the knowledge of the system. Nielsen also used this categorization as one of the concept to define users. The other two concepts of Nielsen are not always be used individually for most studies because, some studies are conducted with specific products, such as a business software for companies and this software requires specific knowledge on the system. Thus, expertise should be defined for participant selection based on the context of the study. In addition, based on the results of the study by Faulkner and Wick (2005), the users divided according to general computer knowledge, did not show significant differences in results.

Beside these approaches, Gerardo (2007) also offers another concept. According to this concept; if there is no expert user, with a short training session before the study, a participant can be called an expert.

As a summary, in the context of this thesis, expert is defined with more than one year of continuous experience on a studied system while the novice is defined has no or little experience with the studied system. In addition, all users will be selected with a general knowledge and experience with touchpad screens.

2.4.2.2 Previous comparative studies on the effects of user expertise

In literature, there are several studies focusing on comparing novice and expert users in usability tests. However, most of these studies focused on the comparison of performance metrics (time on task, success rate, subjective satisfaction… etc.). In formative studies, there are few studies conducted to compare these user groups with the quality outputs (usability problems; number, severity, variety of types) of usability testing.

The previous studies about user expertise were analyzed to find out what differences and effects that the researchers and designers come up with the evaluation of the design. However, the object of the study, i.e. the interface or the product tested, show differences, the approaches, methods and results were mostly parallel.

The first two studies presented below focused on the comparison of performance metrics.

Dillon & Song (1997) compared the novice and expert users on both graphical and textual based search interfaces for a university database on art resources works. They prepared one textual and one graphical based prototype. Through testing of two prototypes with 24 participants divided into 4 groups, they compared task completion time, search performance (accuracy), navigation style (number of points visited in the interface) and responses to a post-task interview. They found that, novices spent significantly more (almost twice as long on average) time than experts. Beside this users spent less time (novice user - %12 and expert user - %15) on graphical interface rather than textual ones but the effect of interface style was not significant. Independently of the interface, experts were better than novices on search performance and novices visited more paths, but these results were not significant.

Similar results were obtained in the study of Ziefle (2002), but setting was little bit different. Ziefle conducted a study to compare three mobile phone interfaces different in terms of menu complexity (depth and breadth). She recorded the performances of 60 participants in total; 20 for each prototype, in all these three interfaces. The performance measures taken were efficiency (completion time and number of detour steps), effectiveness (success rate), perceived ease of use (4-scale rating) and learnability. In order to measure learnability, same tasks were conducted twice with a time period in between tasks. They found that, expert users were better to complete tasks than novice users in all interfaces. In addition, significant interaction was found between expertise and complexity for learnability (novice>experts) and highest learnability was found with the medium complex interface. Average time for efficiency was calculated and results showed that novices spent more time rather than experts. Detour steps also observed for efficiency and it was found that, novices made almost double more steps than experts did.

There are some studies aimed to differences between novice and expert users on uncovered usability problems.

Faulkner and Wick (2005) conducted a cross-user usability test on a web-based employee timesheet application to compare three different user groups to analyze the problems (deviations form defined paths) with number and type indicators. They categorized users according to experience levels with computers in general and the application they were testing. According to this classification the authors grouped 60 participants as; novice-novice, expert-novice and expert- expert. In the first group-

novice-novice, the users had very little experience with computer and had never used the application before. The second group expert-novice included users who had a general computer experience and had no prior experience with the tested application. The users in the last group defined as expert-expert who had more than one year experience on general computer usage and currently use the tested application. First of all, they prepared an ideal path for the system and for each point (45) of this path; they counted all the wrong actions as deviations for each user group. They compared the deviations for all these 45 points and user groups and tried to find differences. They found that, in general, novice-novice users deviate more than other user groups. According to results, there were no significant differences between novice-novice group and expert-novice group while the differences between expert- expert group with other two groups were significant. Overall results showed that, the general computer knowledge affected only some of the determined deviations.

Gerardo (2007) aimed to investigate the effectiveness of using novice and expert users in usability test. He conducted a usability test on redesigned ERP system (business software). He tried to find whether novice and expert users find the same type of problems and how the total number of uncovered usability problems between these groups differs. In this study, 12 expert (have experience with the system) and 12 novice (have no experience with the system) users were used. He compared success rate (task completion), time on task, task difficulty, the number of problems and the type of problems. The results of the study showed that, novice users found more problems than experts did. Although, novice users found same types of problems as experts, novices also revealed additional problems that experts never experienced. The success rates of experts were better than novice group. In addition, novices were experienced more difficulties than experts with the same task and spent significantly more time in total. The researcher concluded that the novice users should be included in usability tests on redesigned systems for effective usability test results.

As a summary, experts are better than novices when the performance metrics (success rate, completion time) are measured in usability tests. On the contrary, novices show better performance to detect more usability problems. In these reviewed studies, types of problems were not defined in detail and no significant

differences were observed with the type of problems between novice and expert groups. None of the studies were measured the severity rates of the problems.

2.4.3 Interaction between prototype fidelity and user expertise

All the studies that were mentioned previously were focusing on the prototype fidelity and the user expertise independently. Beside these studies, there is one study that includes the interaction effect of these two factors. Sauer et al. (2010) studied on the" floor scrubber" and in that study, primary functions; navigation, cleaning and maintenance and system monitoring were tested with 48 participants. They used in their research user expertise (novice and expert) and prototype fidelity (paper, 3D mock-up and fully operational appliance) as independent variables. All these three prototypes are presented in Figure 2.7.

Figure 2.7 : Prototypes of floor scrubber: (a) high-fidelity, (b) medium-fidelity, and (c) low-fidelity. (Sauer et al., 2010)

The measures used in this experimental study were the number of usability problems, the severity of the problems that were identified, types of the problems, performance data (task completion time, water consumption and achieved cleanness), controls settings and system intervention and subjective user ratings. In this study, participants were asked to report usability problems they had experienced with a semi-structured interview following the experimental session. They found that expert users discovered significantly more usability problems than novice ones. The difference between them was larger for the low-fidelity prototypes than for the fully-operational appliance but the difference was not significant according to statistical analysis. Novice users found significantly more severe problems than experts because; they were unfamiliar with the product and the tasks, they more effective in identifying important problems. The severity ratio was higher with 3D mock-up

prototype while the severity ratio was almost same with other two and the relation between expertise and fidelity is significant. The effect of user expertise on aesthetic judgment is significant (novices found the application more appealing than experts). But there was no relation between fidelity and aesthetic judgment. This study reported that, there are differences between prototypes and user groups with regard to the mean number of usability problems in each category. Performance measurements were only calculated with fully interactive (high fidelity) prototype and found no significant differences between user groups. They did not conclude that, one factor; neither fidelity nor expertise is superior to the other. According to them, the purpose of the usability test such as; max usability or most severe problem identification can lead designers and researchers using either novice or expert user; low or high fidelity prototypes.

Sauer et al. (2010) conducted this study with a physical product. The tasks were also very simple, so that small amount of actions can be done with this product. The comments of users were the only source of usability problems and because novices were unfamiliar with the product, they might comment less than experts. In a different way with Sauer et al. (2010), beside the comments of the users, the other main problem source is observation in this thesis. In addition, the product that is used in this thesis is a digital interface and have more complex interface to do multifarious tasks. With this thesis, it is expected to fill this gap in the literature.

3. DESIGN AND CONDUCT OF THE RESEARCH

This section includes the methods that were used for designing and conducting process of the usability test on portable navigation device for cars. The section starts with the purpose of the study, a brief description of usability test and the methods used to gather and analyze the data during this project described in detail. The test object, the tasks, prototypes and other test materials, participants and the test environment and the test procedure are also described in following parts.

3.1 Purpose

The main purpose of this study is to evaluate whether and how prototype fidelity and user expertise influence the outputs of usability testing especially uncovered usability problems.

All sessions were recorded to analyze the reactions of participants in detail. Retrospective think aloud method was used during the tests between each task to gather more data from users about their experience with the interface. By doing this evaluation, usability problems were analyzed in number, severity and variety of types. After that, it was possible to remark which user groups took part actively in which kind of prototypes to provide more data. In addition, the performance data (time on task and success rate) was analyzed to see the differences between both user groups.

3.2 Usability Test

Usability tests are more practical to uncover usability problems provided by the realistic experience rather than imagined one prior to launching a product. It provides researchers direct information about the way of using the system and helps to find some unpredictable problems while studying with users that they cannot discover during the evaluation.

In this thesis, usability testing was used representing the real users as participants and conducting real tasks to evaluate what will happen when the product gets to the real users. To simulate the realistic experience and let participant to complete the task without any interruption "Retrospective Think Aloud Protocol" and "Performance measurements" were used to gather data to achieve the desired goals.

For comparative analysis, "Performance measurements" provides results to understand if the study could achieve the usability goals (Ivory, 2001). In current study, the data was collected with asking the participants to complete the determined tasks. While the participant doing tasks the test session was recorded for further analysis. Performance Measurement method was used to get actual data with analyzing these recordings about how the participant completed the task and how long it took.

The big amount of the data is collected to understand the way that users think to complete an action with the "Thinking aloud Protocol". With this method, participants are asked to speak loudly their thoughts, feelings, and opinions continuously during a usability test while they are performing tasks (Ivory, 2001). Beside the advantage of this verbalizing to get the answer what the problem is, why and how it occurs, keeping participants to talk can be sometimes hard and it can also interrupt to perform a task (Umar and Tatari, 2008). There is another variation of Think aloud protocol named as "Retrospective Think Aloud Protocol". With this method, participants are required to speak loudly their thoughts, feelings, and opinions continuously during a usability test after the task has been completed (Gray and Wardle, 2013). The common technique for this method is to record the whole session and after the current test or each task, while watching the recorded session together with participants, asking them to verbalize every action that they made while completing the task. On the other hand, the test moderator can determine the important parts of the prior task and can only ask participants to report their thoughts that they remember about these parts before the following task (Eger et al., 2007).

In this thesis, Retrospective Think Aloud Protocol was used to gather more information about participants' actions and reasons of their behaviors, by repeating the required parts of tasks. After each task with this method, participants were asked to explain their thoughts about the problems that they uncover and comments that they made in prior task. The moderator also observed participants while completing

tasks. All these data were analyzed by the researcher including her observation to formulate usability problems.

3.3 Test Materials

It is needed to prepare sets of materials to conduct a usability test. First of all a test object was decided and tasks were created depend on the functions of the test object. The product itself was used as a high fidelity prototype and before each test session the settings of the device were changed according to given tasks. A paper prototype was built according to task phases. Beside these, the required materials for recording the test sessions were supplied. These materials are introduced in detail in the following sections.

3.3.1 Test object: portable navigation device

This thesis was written in the Department of Industrial Product Design. Products are changing into digital versions with the technological advancements. This change creates new products with digital interfaces (less physical specifications). A Portable Navigation Device (PND) was chosen as the test object. This product has digital and more complex interface to do multifarious tasks. With these specifications, this study differs from the current literature. In addition, the other advantage of the study with PND as a test object was to find novice and experts users easily. In addition, because it is portable, the test sessions would be conducted in different locations.

This device offers user; route navigation, tourist information; point of interest (POI), emergency services, information about the driving rules in lots of countries, preference options for personal usage…etc. There are different kinds of applications of PND for vehicles such as; car, motorcycle, camper & caravan and trucks. These devices have software inside and gather combination of data from different satellites to provide reliable results.

There are lots of brands producing PND for vehicles while some of them also have mobile navigation applications. In this thesis, TomTom XXL Classic Series was used. The main purpose of the thesis is not to evaluate how usable or not the product of this brand. It was aimed to compare fidelity and expertise influence on uncovered usability problems. The possibility to find more usability problems to gather more data into this study would be high with the use of old products. Therefore, it was

decided to use a product which had an old version (9.061) of the software (lastly updated on September 15, 2010). The participants also informed about the purpose of the study and the product that they used an old version to protect the company's rights. The information was given to the participants after the test not to affect their behaviors.

3.3.2 Tasks

After selection of the test object, one of a current user was observed during the use of the product and what kind of problems he had was reported. The method is also mentioned by Nielsen and Norman Group (2014) as the most effective method to understand the product. Then, a couple of current users were asked about the most frequently used functions. Dumas and Redish (1999) indicated that it is important to define the tasks which "probe potential usability problems". Thus, the interface of the device was evaluated to find the problematic parts. It is also important to verbalize task scenarios which are more suitable for usability testing. To gather the best results from the usability studies, Nielsen and Norman group offers 3 task-writing tips; make the task realistic; actionable and avoid clues and describe the steps (2014).

The PND offers lots of functionalities especially about route navigation. The preferred destination can be inserted by writing the complete address, choosing from the favorite addresses or recent destinations, searching from the specific pinpoints (POI)…etc. For this study, only the most frequently used route navigation tasks and preference options were applied. These functions were; navigating to home and updating the home address; adding favorite address; inserting sub-route into the current route; using quick menu; reading the information and instructions on the main (map) screen; setting a safety warning for the speed limit. In this study, it was aimed to simulate the interaction mostly before to start driving. The tasks also include some duties that can be done while driving.

According to these pre evaluation and reviewed guidelines, a set of test scenarios included 8 tasks were prepared. After a pilot test, tasks were reformulated into 6 test scenarios. In general, all scenarios included the missions; to evaluate finding the right menu, understanding and using the interactive components, finding information

in the single page, understanding the menu titles and icons; some scenarios also required responding the system messages.

First three tasks were related to route planning. First task involved route navigation, updating a home address; second task involved adding the work address as a favorite; third task involved navigation to a previously added favorite address (work address) and adding a sub-route (POI) into this route. Last three tasks were related to preference settings and read the information on the screen. Fourth task involved to create a quick menu with desired functions; fifth task involved to read the main screen (map) to explain the graphics and information on the screen and to edit the desired information; sixth task involved the setting a preference to warn driver when he/she drives too faster than allowed.

Some scenarios were formulated with a relation between each other. For example, the favorite address that was navigated in the third scenario was created in the second scenario. And the function "Use night/ day colors" added in the quick menu during the fourth scenario and it was asked to use this function through the quick menu in the fifth scenario. The brief explanation for the realistic situation for scenarios also added into the speech of the moderator.

After each task, users were asked to say their thoughts about the visited screens and the problems that they had on each step. This additional task provides to gather additional data from users about their experiences. The whole set of detailed scenarios of this study was added in both English and Turkish in Appendix A.

3.3.3 Prototypes

The purpose of the study was to compare two prototypes with different fidelity. In this study, the product itself was used as a high fidelity prototype and the low fidelity paper prototype was prepared based on the specifications of the product interface. The similarity between the prototypes was important. Thus, all the required pages for completing the tasks were prepared for paper prototype (low fidelity) including the same page layout, elements and wordings as the tested product. The paper prototypes did not include color and the images were less clear and representative than high fidelity prototype. In addition, for some tasks, users would tend to follow alternative routes, thus some extra pages were prepared for paper prototype.

McCurdy et al. (2006) offered five dimensions to define a prototype beside the definition of low- or high fidelity. These dimensions are; visual refinement (look of the prototype), breadth of functionality (scope of features), depth of functionality (fidelity of functions), richness of interactivity (similarity of interaction) and richness of data model (actual data). According to these dimensions, the high fidelity prototype (interactive) differs from the low fidelity prototype (paper) mostly in the dimensions "Richness of interactivity" in this study. In addition, because of not using color and using less clear images, these prototypes were also differed from each other with the dimension "visual refinement". Figure 3.1 presents sample pages of the low (paper) and high (the device itself) prototypes used in this study. The detailed features of these prototypes are explained in the following sections and screen samples of these two prototypes are presented in Appendix B

Figure 3.1 : Sample screens of low (paper) and high (the device itself) prototypes

3.3.3.1 High fidelity prototype (the device itself)

As it was mentioned above, in this project the product itself was used as a high fidelity prototype. The product, TomTom XXL Classic Series PND has an old version (9.061) of the software (lastly updated on September 15, 2010). The main purpose of the thesis is not to evaluate how usable or not this product of this brand. It

was aimed to compare expertise and fidelity influence on uncovered usability problems.

Some special settings of the interface were done depend on the related task and before the each test session all settings were changed to initial position.

3.3.3.2 Low fidelity prototype (paper)

The paper prototype was prepared as a hand-drawn version of the device itself. It took only two days to produce all desired screens. Each test had its own paper screen set including the whole process of the scenario, thus some screens were copied for each task. These separate sets were prepared due to not to spend time to organize screen sets for next task.

Figure 3.2 presents the materials that were used for paper prototyping; white paper, ruler, pencil, black pen(thin), black marker (thick), cardboard (black), pritt stick, utility knife.

Figure 3.2 : Paper prototype materials

The screen was drawn on the white paper with the same dimensions (8cm*11cm) and was glued on the cardboard that simulates the border of the original device. As the working principle of the original device, all steps of the scenarios were drawn on separate screens including the pop-ups

3.3.4 Other test materials

Usability tests were recorded with a phone camera and special mounting device for phones was needed to hold it over the table. For this reason, a product named

"System-S Universal Flexible Gooseneck Table and Bed Mount for Smartphone" was supplied. With this product, the test setting was practically prepared. Figure 3.3 shows this product.

Figure 3.3 : System-S Universal Flexible Gooseneck Table and Bed Mount for Smartphone

3.4 Participants

For this study, only a small number of users could be tested due to time and resource restrictions. Because of the small number of participants, it was intended to find a homogeneous group of participants and to not include outliers such as participants with no computer and especially any other device with touch screen experience. Nielsen (1993) pointed out in his book that the test will be dominated by the effects of the user's struggle with the interaction devices and techniques, if users are not trained in the use of them (p. 177). Thus, by choosing participants for this study, it was made sure that all of them were familiar with using at least smartphones.

Usability tests were both conducted in Istanbul-Turkey and Ingolstadt-Germany and participants were chosen among Turkish people. Three of the ten expert participants currently live in Turkey and they have used the device in both Europe and Istanbul. Other experts currently live in Germany and they have used the device only in Europe. The participants were selected according to their experience on the navigation device especially on TomTom. A participant with more than one year of continuous TomTom navigation device (from the version 9.061 up to actual version) usage was considered to have a high experience on tested system. Only one expert user was using another car navigation system, which was integrated into his car. In addition, another expert user currently uses another device and the last time that she

used TomTom device was almost six months ago. According to Gerardo (2007), if there is no expert user, with a short training session before the study, a person can be behaved as an expert. For these two expert users, training sessions were conducted to gain experiences on TomTom before the test began. After that, these two participants could attend on usability tests as experts. Only two of the novice users use car and some of the novice users currently use another navigation app on their mobile phones, but they only use searching functions just to be informed where the address is.

Twenty participants were used in total, dividing into four groups with five participants each. Due to time and resource restrictions of the study, it was not possible to include more participants. In section 3.4.1, it will be explained the approaches about sample size in usability testing.

Before the usability tests, participants were asked about their experiences and usage on TomTom and similar products. This information is presented in Figure 3.4. As shown in Figure 3.3, nine out of ten expert users use TomTom devices in their daily life; one of them uses another device, which is integrated into his car. Only two of the novices use car and currently use Yahoo map app, the other three novices use Google map app on their mobile phones but as they reported, they just use searching functions just to be informed where the address is. Five of the novice users have no experience on any navigation system.

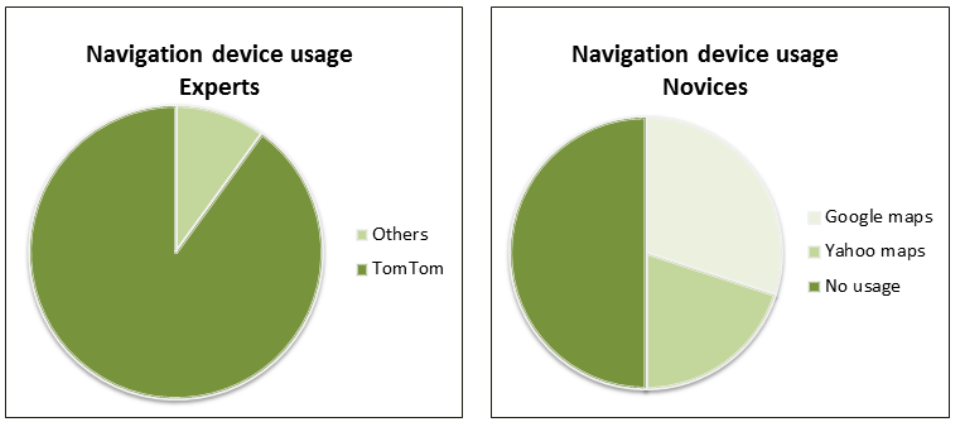

Figure 3.4 : TomTom and similar product usage ratio of user groups

The participants divided into four groups. Group NL refers the novice participants who worked with the paper prototype (low fidelity) while the group NH refers the

participants who worked with the device itself accepted as high fidelity prototype. Beside this, expert participants worked with paper prototype named as group EL and other experts who worked with the device itself named as group EH. Participants' background data is shown below in Table 3.1.

Table 3.1 : Demographic information of participants in usability tests

No	Age	Gender	Education Level	Experience with touchpad usage(years)	Group
1	38	F	High school	3	NL
2	27	M	Bachelor	4	NL
3	29	F	Master	6	NL
4	30	F	Bachelor	5	NL
5	28	F	Bachelor	5	NL
6	29	M	Bachelor	6	EL
7	29	F	Master	5	EL
8	28	M	Bachelor	4	EL
9	33	M	Bachelor	4	EL
10	40	M	Bachelor	3	EL
11	28	F	Bachelor	5	NH
12	33	F	Master	3	NH
13	28	F	Bachelor	4	NH
14	28	F	Master	4	NH
15	28	F	Master	5	NH
16	36	M	High school	3	EH
17	42	M	High school	4	EH
18	28	M	Bachelor	3	EH
19	40	M	High school	3	EH
20	28	M	Bachelor	5	EH

3.4.1 Sample size

"The evaluation of a design element's quality is independent of how many people use it." (Nielsen, 2012b). The main purpose of the usability testing is to evaluate the functionality of the interface and to find if the design elements are easy or difficult to use (Nielsen, 2012b).

In another article, Nielsen mentioned that to get the best results from usability tests, it is recommended to use no more than five users (2000). In his early research with another colleague Landauer, they come up with a formula. Based on this formula, the Figure 3.5 below show us the number of users and the percentage of the usability problems found by that amount of users.

According to this figure, after some amount of users, adding another user to the process provides only small amount of extra information. Some insights from users can be mentioned multiple times. Thus, Nielsen indicates that after fifth user, not much new information is occurred and observing the same findings repeatedly cause the waste of time (2012b).

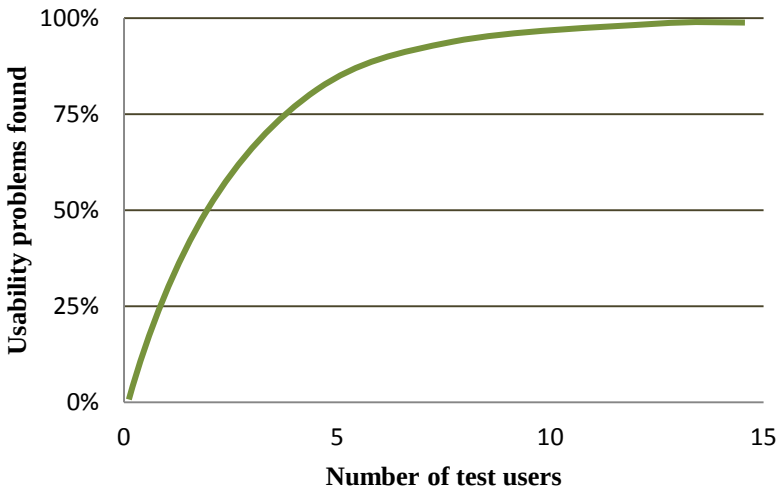

Figure 3.5 : The ratio between number of test users and found usability problems (Nielsen, 2000)

Even if the curve in the figure represents to get the all usability problems, it is necessary to use 15 users. For this situation, he recommends to make multiple tests with 5 users for each to distribute the budget. According to him, the major goal of the usability engineering is to improve the design and not to report all problems of the interface. Thus after each test with 85% of the usability problems by five users, the

following test can be done with more improved interfaces. In literature, this process is named as an iterative design.

Nielsen also defines the number of users in testing with multiple groups. To gain a better outcome from testing with the overlap between observations it is recommended to use 3-4 users from each category if testing two groups of users (2012b).

Sauro (2011) stated that, to find the appropriate number of the users in a usability test, researchers firstly need to identify the type of usability evaluation. As mentioned in section 2.3 these types are; detecting usability problems in an interface, estimating a parameter, making a comparison, comparing to a benchmark. In this thesis the focus is mostly on the usability problems that provided by users.

Similar to Nielsen (2012b), Sauro (2011) mentioned that after each additional user, the percentage of the new usability problems diminishes. He also added that not all the users were affected by each problem. Thus, for example, sometimes a problem can affect just one user in ten. It depends on the influence area of the problem and severity of it.

3.5 Test Environment

The usability tests took place at the participant's own working area or home. The working area has a table to mount the video recording device (mobile phone). In Figure 3.6, the layout of sample test environment is presented.

The test users sat close to the moderator. Thus, moderator was able to see what the participant is doing to run the task steps. Beside this, this seating makes participants feel comfortable. Figure 3.7 presents the test with high fidelity prototype (the product itself) and Figure 3.8 presents the test with low fidelity prototype (paper).

There was only one moderator and it was not possible to include another observer to the tests. The moderator was setting the study environment, observing and conducting the test session at the same time. For this reason, the test sessions were recorded to watch and analyze later. A mobile phone camera was used as a recorder and a mounting device was attached on the table to hold it. The camera was directed

at either the device or the paper prototype and only the hands of users were recorded while doing tasks.

Figure 3.6 : The layout of usability test.

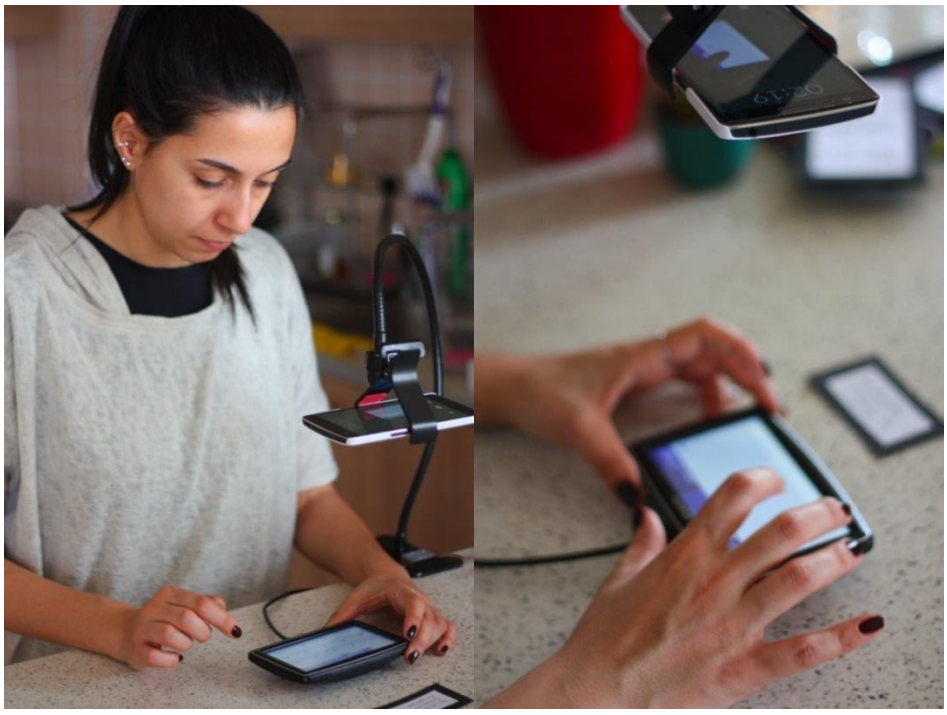

Figure 3.7 : The test session with high fidelity prototype (the device itself)

Figure 3.8 : The test session with low fidelity prototype (paper prototype)

3.6 Procedure

This thesis was written in Ingolstadt by the researcher who also moderated the usability tests and made analysis. Due to the time limit and finding appropriate participants for this study, usability tests were conducted both in Istanbul (Turkey) and Ingolstadt (Germany) and it took two weeks for twenty participants. Two weeks before the usability tests, all participants were informed about the study and meeting times were scheduled. All the demographic data and experience details on navigation systems were collected during the first contact with participants.

First of all, after preparing the first set of scenarios included 8 tasks, a pilot test was conducted. It is important to test the prototype workflow before the first test session to find and fix the potential problems with workflow. Also, with pilot testing, it is possible to reformulate the wording and predict the required time for testing (Schade, 2015). Due to the time limit, 8 task scenarios were reformulated into 6 test scenarios. In addition, the workflows were verified to prepare prototypes.

Almost all tests were done in weekdays and because of a lack of time and working hours of participants; they were visited in their own workplaces or homes. In addition, the test setting was easy to build and it was easier to visit participants instead of inviting them into specific test place.

The time plan was done according to participants' programs. First week, the tests were done with 8 novices and 3 experts in Istanbul. Second week, 2 novices and 7 experts took part in tests in Ingolstadt. The participants were divided into four groups. As mentioned in section 3.3, 5 novice and 5 expert participants worked with the low fidelity prototype (paper) while another 5 novice and 5 expert participants worked with high fidelity prototype (the device itself). Each test comprised with six tasks and took almost 30 minutes. Same tasks were asked to complete for each group.

The same process was followed in all tests. Test scenarios were written on a test guideline for moderator to follow the process. Before the test meeting with high fidelity prototype, the required settings for tasks were done on the device; with low fidelity prototype, paper screens were grouped into task by task. The test setting was also built by moderator which is explained in section 3.5 in detail. Before the test started, brief information about the test purpose, goal and procedure was given to the participants.

Participants worked with the high fidelity prototype (device itself) had opportunity to visit all pages in the device. If a participant wanted to complete the task in another menu, firstly the moderator waited for him/her to realize that they were in the wrong menu. If a participant spent too much time (more than 2 minutes) to find the right page in wrong menu or requested help after more than 5 wrong paths, he/she was informed by the moderator that they were not in the right place to guide participant to complete the task.

Paper prototypes were prepared including the possible necessary pages to complete the tasks. The moderator manipulated the device by changing the appropriate screens according to participant's actions. For some tasks, some participants would tend to follow alternative routes, thus some extra pages were prepared for these tasks. Before the test started, participants were informed about how to work with paper prototype. For example, if a participant wanted to follow the way that was not included in the screen set for that task, the moderator couldn't refreshed the page thus participant realized that he/she was following the wrong way. Then, participant could continue to find the right menu.

Some tasks were formulated including two or three parts. For these tasks, not to confuse the participants, the moderator firstly read the first part and after they completed the first part, second part was read. After each task, participants were asked to explain in detail their experiences on the task process especially on problems that they came up with. In addition, during the test sessions, some participants were willing to discuss about the problems and to ask about how the interaction worked. Due to the test rules, their questions were not replied. Only the questions over reading the text, because of the text quality in low fidelity prototypes, were answered if it was an important item for the next step in a task. After the test sessions, participants were asked to say in general, what kind of problems that they found more important and why. All test sessions were recorded not to lose any valuable data from users and the recordings were analyzed in detail after to collect the usability problems.

In this study, TomTom XXL Classic Series, version 9.061 (the software lastly updated on September 15, 2010) was used as a test object. As it was mentioned in 3.3.1, with using the old version of the product, the possibility to gather more data based on usability problems was higher. Due to protect company's rights, each participant was also informed about the purpose of the study and the product that they used an old version. The information was given to the participants after the test not to influence their behaviors.

4. RESULTS AND ANALYSIS

The data was gathered by the video recordings and notes by the moderator. First, 20 usability test sessions (5 high-experts, 5 low-experts, 5 high-novices, 5 low-novices) were recorded by mobile phone. The average duration for each session was 30 minutes. A large amount of data was collected and analyzed to gather two main outputs; usability problems (number, severity and variety of types) that each user had made and performance data (time on task, success rate).

The following sections include the results of the study in detail. First section describes the method that was used to analyze the usability problems and results; second section includes the method that was used to analyze performance data and results. Last section discusses all results and aims to reveal the goals of the study. The statistics, are presented in Appendix C.

4.1 Analysis of Usability Problems

Informative studies, the main output of the usability evaluation is a set of usability problems. With identifying these problems, designers and developers could fix the problematic area and develop the interface. That is to say, the success of a formative usability test depends on how well the test reveals the usability problems. In order to understand well how these expertise and fidelity settings perform, the number of the uncovered problems, the severity of these problems and the variety of the types of these problems are compared.

In this study, two main sources were used to reveal usability problems. The first source was observing participants and coding their errors (e.g. pushing the wrong button), wrong actions (e.g. to follow the wrong path), indications of confusion (e.g. not being able to see the button or understand the information).

The second source comes from the users themselves; whenever they say something that points to problem also was coded as a problem. In this case, misinterpretations (e.g. to interpret the information in different meaning), confusions (e.g. uncertainty

to continue the action and not being able to understand the information) and other comments (e.g. commenting on lack of information) were coded as problems.

Firstly, all the revealed problems were written down in Excel-sheets for each participant and the problems were tagged with the related problem types and subtypes based on the first categorization (six main problem types were mentioned in section 2.3.1.3.) using post-its. Figure 4.1 presents the post-it categorization of problems. The amount of problems was determined for each group and then the relation was calculated to the total number of problems found to compare participant groups and prototypes.

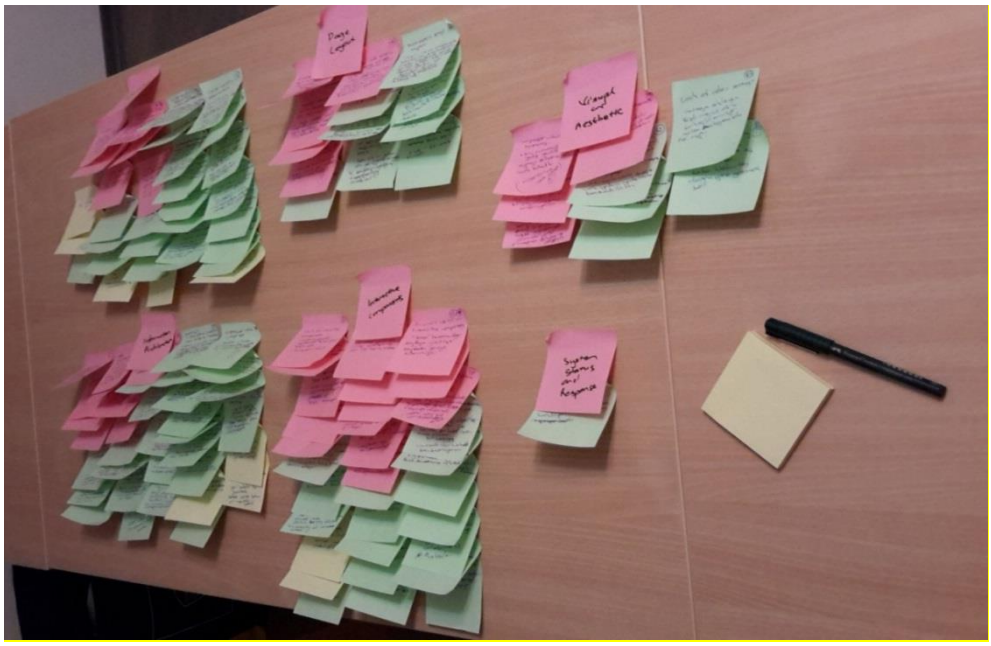

Figure 4.1 : Usability problem categorization with post-it

Secondly, the sets of usability problems listed in detail with attached screenshots if it was needed and two usability experts rated the severity of problems independently. It was already mentioned in the literature section that, it is better to use more than one evaluator to categorize and define the severity ratings of the usability problems for more reliable and valuable results. These raters rated the problems according to determined three severity scales: low severe (minimal usability problem, doesn't prevent user to complete task, can be fixed easily), medium severe (moderate usability problem, impresses user to complete task, takes time to be fixed) and high severe (major usability problem, mostly cause task failure, important to be fixed unless take more time before the product released). The detailed guideline for

severity rating is attached in Appendix D. The number of participants that encountered a problem also considered to rate the severity.

Lastly, based on the context of discovered problems, the reformulated categorization of usability problems are; content (e.g. unclear expression/wording), use flow (e.g. inappropriate number of task steps), page layout (e.g. improper functional grouping and positioning in the page), menu categorization (e.g. unclear menu structure), interactive components (e.g. unclear interactive components), system status and response (e.g. inadequate feedback on where user is in the site) and aesthetic- visual (e.g. inappropriate color usage). In section 4.1.3.1, the detailed version of problem types is explained with the subtypes and definitions of the problems.

Following sections describe the results of the usability problem analysis of the present study.

4.1.1 Results on Number of Problems

The problems revealed by each user counted simply and in total 471 usability problems were identified. After calculating the overall number of problems discovered by each participant "Two-way ANOVA" was used to analyze the differences and relation between novice and expert groups and high and low fidelity prototypes. Table 4.1 presents the mean values and standard deviations for each group and prototype.

Table 4.1 : Mean number of usability problems reported by each user as a function of levels of expertise and prototype fidelity

Prototype	Participants	Mean	Std. Deviation	N
Low Fidelity	Novice	27	4,85	5
	Expert	21,40	5,32	5
	Total	24,20	5,63	10
High Fidelity	Novice	24,60	3,71	5
	Expert	21,20	2,28	5
	Total	22,90	3,41	10
Total	Novice	25,80	4,26	10
	Expert	21,30	3,86	10

When the total usability problems were counted distinctly 102 usability problems were reported (27% by novices, 12% by experts, 61% by both user groups; 27% with low fidelity prototype, 15% wit high fidelity prototype, 58% with both prototypes).

4.1.1.1 Results regarding the effects of prototype fidelity

The difference between low and high fidelity prototypes is not significant but according to average numbers, users discovered more problems with low fidelity prototype (242 vs. 229). Experts found equal amount of problems on both prototypes (107).

4.1.1.2 Results regarding the effects of user expertise

The results showed that, there is a significant relation between expertise and the number of problems. In this study, novice users found significantly (F= 5,720; df= 1, 16; p < 0, 05) more problems than experts (258 vs. 213).

4.1.2 Results on severity of problems

After the first severity rating, the agreement ratio between the raters was 72%. In the literature, this result called as substantial (Landis and Koch, 1977). The differences in severity ratings were discussed by two raters in aother session until they agreed on. This time, the agreement ratio was 92% and this result called as almost perfect agreement (Landis and Koch, 1977) in the literature. Table 4.2 presents the frequencies of ratings for two judges.

Table 4.2 : Frequencies of ratings for two judges

		Rater 1			
		1	**2**	**3**	**Total**
Rater 2	**1**	26	3	0	29
	2	0	42	2	44
	3	0	3	26	29
	Total	26	48	28	102

Table 4.2. shows the agreement or disagreement between raters. 26 problems were rated as *high severe* by two raters, 2 problems were rated *high severe* by rater 1 while rater 2 rated as *medium severe.*

According to second session, the raters were not agreed on only 8 problems (shown in Table 4.2.) and for these problems, based on their scales, the average severity points were given to calculate. *High-severe* problems were rated with three points, *medium-severe* problems were rated with two points and *low-severe* problems were rated with one point. For example, a problem was rated *medium severe* by first rater and *high severe* by secon rater. The average severity point for this problem is; (3+2)/2=1.5. After multiplying the severity points of discovered problems and dividing the total into the amount of problems found by each participant, the severity score could be found by participant. By doing this division, the effect of the amount of problems was excluded.

After calculating the overall severity scores of each participant "Two-way ANOVA" was used to analyze the differences and relations between novice and expert groups and high and low fidelity prototypes. The data is presented in Table 4.3

Table 4.3 : Severity ratings by usability experts (1: low; 2: medium; 3 high)

Prototype	Participants	Mean	Std. Deviation	N
Low Fidelity	Novice	2,32	0,06	5
	Expert	2,36	0,17	5
	Total	2,34	0,12	10
High Fidelity	Novice	2,37	0,13	5
	Expert	2,48	0,13	5
	Total	2,43	0,14	10
Total	Novice	2,34	0,10	10
	Expert	2,42	0,16	10

4.1.2.1 Results regarding the effects of prototype fidelity

The difference between low and high fidelity prototypes is not significant but according to average numbers, users found more severe problems on high fidelity prototypes.

4.1.2.2 Results regarding the effects of user expertise

The difference between novice and expert users is not significant but according to average numbers, expert users found more severe problems.

4.1.3 Results on variety of problem types

4.1.3.1 Refinement of variety of problem types

After usability tests, based on the context of discovered problems, the categorization of usability problems were reformulated and divided into seven types and related subtypes. The final categorization and definitions of the problems are presented in Table 4.4.

Table 4.4 : Usability problem categories

Type	Subtype	Definition
Content	Complexity in content	Too much information on a single page Unnecessary information
	Unclear information	Lack of conciseness, translation Unclear expressions, translations , Unclear wording, terms and abbreviations
	Unclear iconography	Unclear meaning of the icon Improper icon usage
	Inconsistent information	Duplicated or contradictory information Neither content nor action related items
	Inconstancy in information quality	Out-of-date and untrustworthy information
	Lack of information	Lack of detail and explanation
Use flow	Inappropriate number of task steps	Too much depth and length in task steps
	Unclear and inefficient task steps	Lack of shortcuts, Uncertainty of how to continue task Unexpected navigation Improper order of task steps
	Improper content of task steps	Irrelevant task steps in a single task Irrelevant content of task steps
Page Layout	Improper functional grouping and positioning in the page	Uncertainty of relation between items because of their improper location Not to notice an information because of location in a page
	Inconsistent page layout	Inconsistency between content related pages

Table 4.4 (continued) : Usability problem categories

Type	Subtype	Definition
Menu categorization	Unclear menu structure	Uncertainty to find appropriate path
	Unstructured and nonhierarchical ordered menu and submenu items	Neither content nor action related categorized menu items Unstructured menu items Nonhierarchical ordered menu items
Interactive Components	Unclear interactive components	Incognizable interactive elements. Not clearly identified interactive and non-interactive components Uncertainty of component interaction (Insufficient function of interactive component) Unexpected interactions
	Lack of interactive components	Lack of navigation and/or confirmation button
	Inconsistent or improper usage of interactive components	Inconsistent format regarding to concept Inconsistent usage of component Improper usage of component
	Uncertainty of component interaction	Insufficient function of interactive component Unpredictable interaction
System Status and Response	Lack of feedback about task steps and completion	Unpredicted changes while completing task
	Unnecessary feedbacks during the task	Too much information, warning, confirmation messages
	Unclear feedback	Not to understand the feedback message
	Inadequate feedback on where user is in the site	Lack of information about the path that is followed
Aesthetic and Visual	Inappropriate color usage	Redundant /wrong color usage Lack of color contrast
	Inappropriate text usage	Inconsistency of text size Inconsistency of text font
	Inappropriate image usage	Inconsistency of image size Inappropriate image concept
	Visual complexity	Lack of distance between interactive items Uncertainty of active/ inactive and selected elements
	Improper visual hierarchy	Improper hierarchy between items and layers

4.1.3.2 Variety of problem types

Usability problems were divided into these seven main types. Figure 4.2 illustrates how 102 distinct usability problems associated with seven problem categories. Based on the results, content related; 36 (e.g. unclear information,), use flow related; 15 (e.g. unclear and inefficient task steps), page layout related; 6 (e.g. improper functional grouping and positioning in the page), menu categorization related; 14 (e.g. unclear menu structure), interactive components related; 20 (not clearly identified interactive and non-interactive components), system status and response related; 2 (e.g. inadequate feedback where user is in the system) and aesthetic-visual related; 9 (e.g. inappropriate color usage) were found.

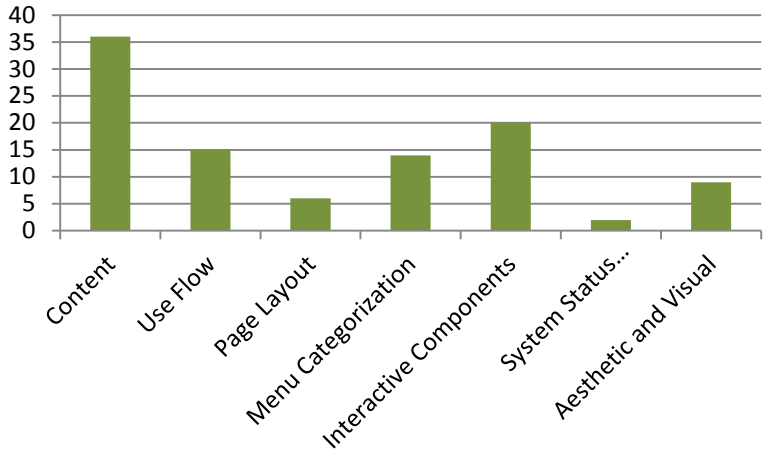

Figure 4.2 : Distinct usability problems by categories

Content related problems

The most frequently observed problem type was *content*. In total, 36 distinct *content* related problems were discovered and the most common problems were based on not clearly identified information (expressions, wordings… etc.), complexity in content, icons and inconsistent information.

Example 1- Unclear information-unclear expression (High severe)

- Participants couldn't predict the content of the menu. The title of the menu did not give the right information what had inside because it was translated improperly.

  The menu "Safety preferences" includes some settings about warnings related to speed limit. The title was translated in Turkish with the meaning as -Safety lock is activated-"Güvenlik kilidi devrede (TR)". Therefore participants thought that it was related with screen lock, child lock, pin code, route lock… etc.

Example 2- Unclear information-unclear expression (High severe)

- The function of the button that finishes the task and navigates to map screen was understood wrongly because of its expression.

  The word on the button "Done" was translated into "Bitti" in Turkish that means "ended, over, finished" and participants thought that if they had pushed the button, it would automatically have closed the screen and canceled the task.

Example 3- Unclear iconography –improper icon usage (High severe)

- The icon on the button was not clearly understood, thus participants didn't act to push that button to open quick menu

  A left arrow was used on the button (in map screen) that opens the quick menu was understood as "undo function, left turn function or navigating to previous menu or main menu". The quick menu button is presented in Figure 4.3.

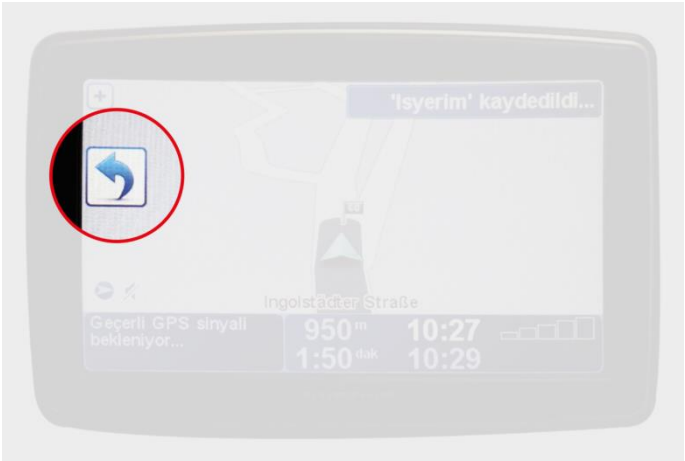

Figure 4.3 : Quick menu button

Example 4- Complexity in content –unnecessary information (High severe)

- Passive items in a single page caused confusion if they were not used in any case.
  The system uses same layout and items inside but, participants were confused with passive items and could not relate with the required action. E.g.: "Home" menu item in "change home location" menu, " Favorite" menu item in "Add as favorite" menu.

Example 5- Inconsistent information –Duplicated or contradictory information (High severe)

- The address that was added as a favorite with a special name still appeared separately in the recent destination list.

  Participants had confusion about seeing the same address both with detailed version and as a favorite with a special name in a list as different addresses.

Interactive components related problems

The second most frequently observed problem type was *interactive components*. In total, 20 distinct *interactive components* related problems discovered and the most common problems were based on unclear interactive components (not clearly identified interactive and non-interactive components, incognizable interactive element), lack of interactive component.

Example 1- Unclear interactive components – not clearly identified interactive and non-interactive components (High severe)

- Novice participants didn't realized that the map screen itself acted as navigation button to reach the main menu.

  Participants tried to find any special button for this action and they pushed "zoom in /out, compass icon, symbol of a vehicle and the information panel below the screen"

Example 2- Unclear interactive components – not clearly identified interactive and non-interactive components (High severe)

- The information panel below the screen was not realized that it was also interactive and when push the right/ left part, a function or another screen was opened. The information panel is presented in Figure 4.4.

  Participants thought that this part was just read-only and gave information about route.

Figure 4.4 : **Information panel**

Example 3- Lack of interactive components (High severe)

- The lack of "back" button made the navigation between pages in a menu difficult.
  Participants wanted to go back but they had to visit all next pages with "next" button to reach the previous page.

Use flow related problems

The third most frequently observed problem type was *use flow.* In total, 15 distinct *use flow* related problems were discovered and the most common problems were based on unclear and inefficient task steps (lack of shortcuts, uncertainty of how to continue task… etc.), inappropriate number of task steps (too much length in task steps).

Example 1- Unclear and inefficient task steps- improper order of task steps (High severe)

- The current flow to select POI made the participants confused. (… > ilgi çekici nokta-POI > şuradan git –navigate from> ilgi çekici noktalar listesi- POI list >…)
  Participants made a comment about the POI category list page should be before the page that they had chosen the location preference. Because in some case there may be no POI that they required in the selected location. Therefore, participants had to turn back to previous steps to complete the required task.

Menu categorization related problems

The fourth most frequently observed problem type was *menu categorization*. In total, 14 distinct menu categorization related problems discovered and the most common problems were based on unclear menu structure.

Example 1- Unclear menu structure (High severe)

- The participants couldn't find the menu to add additional route into the current route.
  The structure of the menu categorization made the participants confused, and they tried to do this task in other menus, which are also related to route preferences and separately located in the system. Participants tried to complete this task via "navigate to, prepare a route, itinerary planning, calculate alternative, map corrections… etc." instead of "find alternative- travel via".

Example 2 -Unstructured and Nonhierarchically ordered menu and submenu items - Neither content nor action related categorized menu items (High severe)

- The current structure of the "change preferences" menu made the participants confused.
  The sub menu items were not grouped neither content nor action related. With this structure it was hard to find to right menu and that problem increased the cognitive load.

Aesthetics & visual related problems

The fifth most frequently observed problem type was *aesthetics & visual*. In total 9 problems discovered and the most problems were based on visual complexity.

Example 1-Visual complexity- Lack of distance between interactive items (High severe)

- The inadequate distance between the lines made to select desired option with “radio button” difficult. Participants made a comment that it could be even harder while driving a car.

Example 2 -Visual complexity- Inconsistency of text size (High severe)

- The text in the toast message popups are not big enough to read

  This message popups appear for couple of seconds and participants commented that it could be even harder to read the text inside especially while driving a car.

Page layout related problems

The sixth most frequently observed problem type was *page layout*. In total, 6 distinct *page layout* related problems discovered and the most common problems were based on improper functional grouping and positioning in the page (uncertainty of relation between items).

Example 1-Improper functional grouping and positioning in the page - Not to notice an information because of location in a page (High severe)

- The toast message popups were not recognized by participants because of their locations were not in the focused area(top-right corner) for this device.

  Participants reported that they couldn’t realize the message popups because with this type of small screens; they mostly focused on the center of the screen.

Only 2 problems were reported related with the type *system status and response*. One of them is inadequate feedback on where user is in the site, the other is related to lack of feedback about task steps and completion. These problems were rated as medium severe.

At last, the “right arrow” in menu pages that navigates to next page was reported as problems related with three different categories and each was calculated separately. First one was in the category “unclear iconography”; Because of the icon shape, the button was understood as “play” button. Second one was related to “inappropriate image usage”, the button was the same size and visually similar with other menu icons. Therefore it was understood as a menu item. Third one was related to the category “improper functional grouping and positioning in the page”. The button was on the list of other menu items and because of the location, it was not clear that the

button itself had a function that affected whole page. The “next button” is presented in Figure 4.5.

Figure 4.5 : “Next” button

The mean number of usability problems identified by users in each category as a function of expertise and prototype fidelity presented in Table 4.5 below “MANOVA” was used to analyze the differences and relation between novice and expert groups and high- and low fidelity prototypes.

4.1.3.3 Results regarding the effects of prototype fidelity

Content

The results showed that, there is a significant relation between fidelity and content related problems. In this study, users found significantly (F= 4,519; df= 1, 16; p<0,05) more problems mostly about “unclear information (expressions, wordings, abbreviations)” and “lack of information” with low fidelity prototype.

Aesthetic & visual

The results showed that, there is a significant relation between fidelity and aesthetic &visual related problems. In this study, users found significantly (F= 17,043; df= 1, 16; p < 0, 05) more problems with high fidelity prototype.

The difference between low and high fidelity prototypes is not significant with the other problem types, but according to average numbers, users discovered more use flow and page layout related problems with low fidelity prototype; menu categorization, interactive components (incognizable interactive elements, uncertainty of interaction and inconsistent or improper usage of interactive

components) and system status & response related problems with high fidelity prototype.

Table 4.5 : Mean number of usability problems from each category reported by each user as a function of levels of expertise and prototype fidelity

Problem Type	Participants	Low fidelity prototype	High fidelity Prototype	Total
Content	**Novice**	10,8	6	8,4
	Expert	7,2	6,4	6,8
	Total	9	6,2	7,6
Use Flow	**Novice**	2,4	2,4	2,4
	Expert	4,6	3,4	4
	Total	3,5	2,9	3,2
Page Layout	**Novice**	0,8	0,8	0,8
	Expert	1	0,2	0,6
	Total	0,9	0,5	0,7
Menu Categorization	**Novice**	6,6	6,2	6,4
	Expert	4,8	6	5,4
	Total	5,7	6,1	5,9
Interactive components	**Novice**	4,8	5,4	5,1
	Expert	2,6	3	2,8
	Total	3,7	4,2	4
System status and response	**Novice**	0,4	0,4	0,4
	Expert	0,2	0,6	0,4
	Total	0,3	0,5	0,4
Aesthetic and visual	**Novice**	1,2	3,4	2,3
	Expert	1	1,6	1,3
	Total	1,1	2,5	1,8

4.1.3.4 Results regarding the effects of user expertise

Use flow

The results showed that, there is a significant relation between expertise and use flow related problems. In this study, expert users found significantly (F= 7,642; df= 1, 16; p<0,05) more problems mostly about “too much depth in task steps” and “improper order of task steps.

Menu categorization

The results showed that, there is a significant relation between expertise and menu categorization related problems. In this study, novice users found significantly (F= 4,762; df= 1, 16; p<0,05) more problems mostly about “unclear menu structure”.

Interactive components

The results showed that, there is a significant relation between expertise and interactive components related problems. In this study, novice users found significantly (F= 10,796; df= 1, 16; p<0,05) more problems mostly about “incognizable interactive elements” and “uncertainty of interaction”.

Aesthetic&visual

The results showed that, there is a significant relation between expertise and aesthetic & visual related problems. In this study, novice users found significantly (F=8,696; df= 1, 16; p<0,05) more problems.

The difference between novice and expert users is not significant with the other problem types, but according to average numbers, novice users discovered more content (unclear information ‘expressions, wordings, abbreviations’ and lack of information) and page layout (improper functional grouping) related problems. In addition, novice and expert users discovered same amount of system status & response (inadequate or lack of feedback) related problems.

Finally, significant interaction was observed between expertise and fidelity with the category “aesthetic and visual” (F= 5,565; df= 1, 16; p < 0, 05).

4.2 Analysis of Performance Data

The two indicators of the performance data; success rate and time on task were calculated in this study. For the analysis of success rate, each task in each session was rated with three scales according to the way of the task completeness. If the participant completed the task directly without an help, it was rated as *success*; if the participant completed the task with indirect guidance of moderator (after 2 minutes waiting or help request from user after more than 5 wrong path), it was rated as *success with help;* if the participant completed the task with the direct guidance (revealing the path that is followed to complete the task) of moderator, it was rated as *failure.* In the current study, *success* was rated with point 2, *success with help* was rated with point 1 and *failure* was rated with point 0. The overall success rate for each participant was calculated by adding up scores for the individual tasks.

The other performance measurement "time on task" was calculated according to following instructions. Start point was the first action of the participant after the moderator finished to reading the task. The end point was the last moment of the user action to complete the asked task. The following sources of interruptions during the sessions were excluded from time on task completion.

- The time for reading and explaining the sub-parts of the task during the individual task session.
- The time for discussion over the task itself, the uncovered problems or not task related subjects.
- The waiting time for the system response (for high fidelity prototype; GPS response and for low fidelity paper prototype; the time for the changing the screens by moderator)

4.2.1 Results on success rate

After calculating the overall success rate for each participant "Two-way ANOVA" was used to analyze the differences and relation between novice and expert groups and high- and lo w fidelity prototypes. The maximum score in each category is 12 (6 tasks, maximum 2 points each, if all were rated as "success"). The data is presented in Table 4.6.

4.2.1.1 Results regarding the effects of prototype fidelity

The results showed that, there is no significant relation between fidelity and success rate but according to average numbers, both user groups were more successful to complete the tasks with low fidelity prototypes.

Table 4.6 : Mean and standard deviation for success rate

Prototype	Participants	Mean	Std. Deviation	N
Low Fidelity	Novice	9,00	1,414	5
	Expert	10,20	1,483	5
	Total	9,60	1,506	10
High Fidelity	Novice	8,20	2,775	5
	Expert	9,80	2,387	5
	Total	9,00	2,582	10
Total	Novice	8,60	2,119	10
	Expert	10,00	1,886	10

4.2.1.2 Results regarding the effects of user expertise

No significant relation between expertise and success rate was found but according to average numbers, expert users were more successful to complete the tasks.

4.2.2 Results on time on task

In this study, total time (in seconds) of six tasks for each participant was calculated manually with excluding the parts written above. To analyze the differences and relation between participant groups and prototypes, “Two-way ANOVA” was used. The data is presented in Table 4.7.

The results show that novice users performed significantly slower than experts ($F= 5,682$; $df= 1,16$; $p < 0,05$) which was expected. There is no significant difference between low and high fidelity prototypes, however it can be noted that the average of the total time on task under low fidelity prototype is higher than high fidelity one.

4.2.2.1 Results regarding the effects of prototype fidelity

The results showed that, there is no significant relation between fidelity and time on task but according to average numbers, both user groups spent less time to complete the tasks with high fidelity prototypes.

Table 4.7 : Mean times and standard deviations (in seconds) for each user group

Prototype	Participants	Mean	Std. Deviation	N
Low Fidelity	Novice	949,80	230,27	5
	Expert	703,20	223,67	5
	Total	826,50	250,38	10
High Fidelity	Novice	854,80	204,35	5
	Expert	684,20	91,40	5
	Total	769,50	174,23	10
Total	Novice	902,30	211,26	10
	Expert	693,70	161,39	10

4.2.2.2 Results regarding the effects of user expertise

There is a significant relation between expertise and time on task. In this study, as it was expected, expert users spent significantly ($F= 5,682$; $df= 1,16$; $p < 0,05$) less time to complete the tasks.

4.3 Discussion

This chapter discusses the results of the study addressing the research questions. As mentioned before, the main goal of this study is to investigate and understand how prototype fidelity and participant expertise influence on the usability test outputs. In this study, two user groups with different level of expertise were used. The purpose was to find out if the novice users are effective as expert users in usability tests to uncover usability problems and how and why these two groups differ in the reported data. In the same way, two prototypes with different fidelity levels were used and the provided data from both low-and-high fidelity prototypes were compared to understand which one is more sufficient in which conditions and with which use

groups. To achieve this goal, the results were analyzed and compared with the previous research into two main subjects; “Usability problems: number, severity and variety of types” and “Performance data: success rate and time on task”

To better discuss and understand the results of this study, the comparative analysis was made with the previous researches based on the research questions. For this evaluation, the results from number of problems, severity of problems, variety of problem types, success rate and time on task were compared between user groups and prototypes in detail.

4.3.1 Effects of prototype fidelity

4.3.1.1 Effects on number of problems

The results showed that, there is no significant relation between fidelity and number of problems but, according to mean numbers, both user groups discovered more problems with low fidelity prototype (242 vs. 229).

In this study, user comments are used to define the usability problems and effects of them as ones of the main sources. With slower process under low fidelity prototypes (because a moderator manipulated the system), users had time to think more over the interface and this result indicated that people made more comments (more comments mean more problems) on the low fidelity prototypes. Mäuselein (2007) also found that, users made more comments on low fidelity prototypes. In addition, both user groups had little or no hesitation while working with low fidelity because there was no possibility to make a real mistake (e.g. data loss, unexpected changes). Thus, they felt more relax to comment.

Some previous studies also supported the effectiveness of low fidelity prototypes to discover more usability problems. Sauer et al. (2010) used in their research three prototypes of a “floor scrubber” with different fidelity (paper as low, 3D mock-up as medium and fully operational appliance as high). They also found that the amount of problems was differing based on fidelity and users discovered more usability problems on low fidelity prototype but the results were not statistically significant. Beside this, Sefelin et al. (2003) conducted two studies using two different systems and for each system they developed computer and paper prototypes with similar functionality. According to results of study 2 the amount of problems was higher on low fidelity prototype. On the contrary, the results of the study 1 showed that, high

fidelity prototype provided more usability problems. With fully interactive high fidelity prototype, users were freer to tour between pages. Therefore, users had chance to explore the system and experiment the different workflows on high fidelity prototype so that, they might discover more problems. Overall result of their study was, there was no significant fidelity effect on the amount of problems. Similarly, some other studies also supported that the fidelity has little or no effects on the amount of problems even if there are some individual problems provided by both prototype groups in later stages (Virzi et al., 1996; Lim et al., 2006) and the early stages (Tam, 2006; Mäuselein, 2007 and Walker et al., 2002) of the design process.

4.3.1.2 Effects on severity of problems

There is no significant effects of the fidelity on severity of problems; high fidelity prototype provide little more severe problems than low fidelity one.

In a similar way, Mäuselein (2007) reported that high fidelity prototypes provide significantly more severe problems. Beside this, some other studies also supported that the fidelity has little or no effects on the severity of problems. Sauer et al. (2010) also found that the severity rating of low and high fidelity prototypes was almost same while the problems provided by medium fidelity prototype were more severe. Similarly, Walker et al. (2002) reported no significant difference.

4.3.1.3 Effects on variety of problem types

Analyzing the problems based on the studied prototypes gives information about which kind of problems are experienced more by which prototypes and it is also possible to indicate why they exist. When the problems were categorized, the results showed that, the amount of problems differed according to types.

Some reviewed studies also found fidelity related differences in results (Magnussen, 2010; Walker et al., 2002). However the problem categories that considered in previous studies are based on the context of the study, Sefelin et al. (2003), Mäuselein (2007), Lim et al. (2006) and Sauer et al. (2010) also reported that low and high fidelity prototypes provide different problems, but with no significant results. On the other hand, other researchers found that both low and high fidelity prototypes revealed similar results based on uncovered usability issues (Tam, 2006; Virzi et al., 1996)

Content

The results showed that, there is a significant relation between fidelity and content related problems; users discovered more problems with low fidelity prototype. The content of the pages are same on both prototypes. Thus, the difference could be due to the usage methodology of prototypes. With slower process under low fidelity prototypes (because a moderator manipulates the system), users could realize if there was a mistake or improper logic in context.

Aesthetic & visual

The results showed that, there is a significant relation between fidelity and aesthetic & visual related problems; users found significantly more problems with high fidelity prototype. High fidelity prototypes require more aesthetic effort, thus users think that, these prototypes are more close to final design and they have higher tendency to comment on these aesthetic specifications. On the contrary for low fidelity prototypes, due to the sketchy appearance, users think that they are unfinished do not consider on these kind of problems.

To our knowledge, the current literature does not provide such a category aesthetic&visual, therefore there is no discussion. Only Sefelin et al. (2003) added that more graphical related comments provided by computer (high) prototypes.

The difference between low and high fidelity prototypes is not significant with the other problem types, but according to average numbers, users discovered more use flow and page layout related problems with low fidelity prototype; menu categorization, interactive components and system status & response related problems with high fidelity prototype. Interactive problems might be observed due to the interactivity problems of touch screen.

4.3.1.4 Effects on performance data (success rate and time on task)

In this study, users spent more time on low fidelity prototype however the success rate was almost same with high fidelity one. However the results were not significant in this study, revealed previous studies reported that, the performance data was better with high fidelity prototypes (Tam, 2006; Mäuselein, 2007). This is due to have some problems to focus on to completing task while waiting for the next screen load with paper prototype (Tam, 2006).

4.3.2 Effects of user expertise

4.3.2.1 Effects on number of problems

The results showed that, there is a significant relation between expertise and the number of problems; novice users found significantly more problems than experts.

The behavior logic of expert users was more action oriented and they already know what they are looking for. Thus, in some cases experts didn't report a problem if they had enough information to complete the task and the process was not influenced by anything. On the contrary, novice users behave after comprehensive investigation and tried to decide the required way during the session. In addition, novices have little or no knowledge about the system, due to unfamiliarity, they made more comments (more comments mean more problems) rather than experts.

Magnussen also came up with the similar result in his study that, people made more comments on the functions and interactions which they unfamiliar with (2010). On the contrary Sauer et al. (2010) reported that expert users discovered more usability problems than novice ones and the difference between them was larger for the low-fidelity prototypes than for the fully-operational appliance but the difference was not significant according to statistical analysis. This difference could be due to the source of problems in that study. They gathered all data to identify problems from the post questions. The studied object required physical effort and experts could comment more based on their previous experiences. Some researchers also found that, there are differences with the uncovered problems but not statistically significant (Faulkner and Wick, 2005; Sauer et al., 2010). In addition, the results of the study by Gerardo (2007) showed that same types of the problems uncovered by novice and experts.

4.3.2.2 Effects on severity of problems

The difference between novice and expert users is not significant but according to average numbers, expert users found more severe problems.

This result has no statistical power and there is only one study that evaluated the severity of discovered problems by user groups to compare. Sauer et al. (2010), reported that, the severity ratio among user groups was almost same with low and high fidelity prototypes.

4.3.2.3 Effects on variety of problem types

Analyzing the problems based on the user groups gives information about which kind of problems are experienced more by which groups and it is also possible to indicate why they exist. When the problems were categorized, the results showed that, the amount of problems differed according to types.

Use flow

The results showed that, there is a significant relation between expertise and use flow related problems; expert users found significantly more problems.

Expert users already know the general concept of the system and had idea about the task steps. Thus, they could comment more on this kind of problems also by considering their previous experiences.

Menu categorization

There is a significant relation between expertise and menu categorization related problems; novice users found significantly more problems.

Experts had knowledge about the content and general structure of the system and they could more active to find unnecessary or improperly located items more easily than novices (e.g. the structure of the "change preferences" menu made participants confused. Because, the sub menu items were not grouped neither content nor action related, this structure increased the cognitive load).

Interactive components

The results showed that, there is a significant relation between expertise and interactive components related problems; novice users found significantly more problems.

Novice users transferred previous experiences with other interfaces into tested one to complete the task if the current interactive buttons were not realized. For instance, one novice user swiped for zoom in/ out on map screen, changing the screen in menu and scrolling in a list.

Aesthetic&visual

The results showed that, there is a significant relation between expertise and aesthetic & visual related problems; novice users found significantly more problems.

The difference between novice and expert users is not significant with the other problem types, but according to average numbers, novice users discovered more content and page layout related problems. In addition, novice and expert users discovered same amount of system status & response related problems.

While the connotation of the word was mostly sufficient for experts to understand the statement, novice users focused on more details such as the grammar used to formulate the words and think more on logical connections with action. The language of studied interface was not originally Turkish. Thus, if the translation does not provide the exact meaning of terms and expressions, users can understand the statement wrong. In general, novices had confusion to understand some terms, expressions and abbreviations and this confusion provided problems. In addition, Turkish words are derived with some special endings and sometimes the meaning of the statement can be completely different if the ending is improper.

4.3.2.4 Effects on performance data (success rate and time on task)

In this study, novice users spent significantly more time than experts. Expert users were more action oriented and they already know what they are looking for and the time that they spend is more important for them. On the other hand, novice users behave after comprehensive investigation and try to decide the required way during the session and for novices. Thus, it is more important whether the system is easy to understand or not.

This result also supported by the literature (Dillon&Song, 1997; Ziefle, 2002, Faulkner and Wick, 2005; Gerardo, 2007). Dillon and Song (1997) mentioned that novices had a tendency to visit more paths than experts. On the other hand, the results of the study by Sauer et al. (2010) showed that the difference with task completion times is not significant between novice and experts.

After calculating the overall success rate for each participant, it is found that, expert users were more successful (not significant) than novices to complete tasks in total**.** In this study, menu categorization problems and unclear expressions in titles affected task completion success. This result also supported by other studies (Ziefle, 2002; Faulkner and Wick, 2005).

4.3.3 Effects of both prototype fidelity and user expertise

The only significant interaction was observed between expertise and fidelity with the problem category “aesthetic and visual”. Sauer et al. (2010) was the only approachable study that includes the user expertise as one of the key factor that affects usability tests with different fidelity prototypes. They observed significant interaction between expertise and prototype fidelity with the severity of problems. The both results represent different contents, therefore, there is no discussion.

5. CONCLUSIONS AND RECOMMENDATIONS

The goal of this thesis is to contribute to the literature by looking at the influence of prototype fidelity and user expertise on usability testing outputs of a digital interface and interaction between these factors if any.

The main purpose of the usability tests is to gather high valuable and reliable data from real users. Users as subjects and prototypes as objects are the key elements of usability testing. In this thesis, usability testing was used to represent the real users as participants and conduct real tasks to evaluate what would happen when the product gets to the real users. To simulate the realistic experience and let participant to complete the task without any interruption beside the observation, "Retrospective Think Aloud Protocol" and "Performance measurements" methods were used to gather data to achieve the desired goals.

For this study, twenty participants have been used in total, dividing four groups with five participants each. 5 novice and 5 expert participants worked with the low fidelity (paper) prototype while another 5 novice and 5 expert participants worked with the high fidelity (device itself) prototype. Each test comprised with six tasks and took 30 minutes in average. Same tasks were asked to complete for each group.

All sessions were recorded to analyze the reactions of participants in detail and retrospective think aloud method was used during the tests to gather more verbally data from users about their experience with the interface. By doing this evaluation, the number of the usability problems, the severity of the usability problems that are identified and the variety of problem types were defined as the quality indicators. After that, it was possible to remark which user groups took part actively in which kind of prototypes to provide more data. In addition, the performance data (time on task and success rate) was analyzed to see the differences between both user groups. However, the sample size of five participants per each group has little statistical

power, this study provided valuable data to predict and understand the influence of expertise and fidelity on usability outputs.

The results of the study showed that;

- Users found more problems with low fidelity prototype.
- Novice users found significantly more usability problems than experts in total.
- Expert users found more severe problems under both prototypes, but the results were not significant.
- When the usability problems were counted distinctly, 102 usability problems were reported, of which 61% were discovered by participants of both groups, 27% by novices only and 12% by experts only. Beside this, 58% of the total numbers of distinct problems were discovered with both prototypes, 27% with low fidelity prototype and 15% with high fidelity prototype. In addition,

Usability problems that were uncovered from the usability tests were analyzed according to context and divided into seven main groups (content, use flow, page layout, menu categorization, interactive components, system status& response and aesthetic & visual) and related subgroups.

- Users worked with high fidelity prototype significantly more concerned with aesthetic & visual related problems. In addition, with the types “menu categorization”, “interactive components” and “system status and response”, the average numbers were higher than low fidelity prototype.
- On the other hand, users worked with low fidelity prototype discovered significantly more content related problems. Beside this, the average numbers of use flow and page layout problems are higher on low fidelity prototype.
- Most usability problems reported with “content” that also discovered more by novice users
- Novices revealed significantly more problems than experts in the types “menu categorization”, “interactive components” and “aesthetic & visual”.
- Experts revealed more problems with the category “use flow”.
- Finally, significant interaction was observed between expertise and fidelity with the category “aesthetic and visual”.

In this thesis, time on task and success rate were reported as performance data of participants. For this analysis, each participant and task were analyzed separately. According to results;

- Both novice users and expert users completed tasks more successfully with low fidelity prototype but this result is not significant.
- Expert users were more successful to complete the tasks in total but there was no significant difference between groups.

The results were significant between the groups based on task completion times;

- Both novice and expert groups spent less time to finish the tasks with high fidelity prototype. But, the prototype fidelity did not significantly influence the completion time.
- Novice users completed tasks significantly slower than experts.

5.1 Final Remarks

Do we really need fully interactive and visually perfect prototypes to understand the system is usable or not? Do experts always perform well and provide all data we needed or is there any uncovered data that we can only get from novices? It was expected to contribute to the literature by looking at the influence of prototype fidelity and user expertise on usability testing outputs of a digital interface and interaction between these factors if any. The main contribution with these revelations will be providing knowhow for those who want to design specific usability tests. In other words, the study was aiming to providing a guideline for usability testing, regarding the issues of prototype fidelity and participant selection.

With this study, the main conclusion is the problem type list that was identified in detail to address the discovered usability problems in tests. Beside, according to results, a sample guideline was indicated;

If the main aim is to discover more usability problems,

- Low fidelity prototypes can be effective as high fidelity ones at the later stages of the design process. Both user groups have little or no hesitation while working with low fidelity because there is no possibility to make a real

mistake (e.g. data loss, unexpected changes). Thus, they feel more relax to comment.

- Novice users are better to find more problems.

If the main aim is to evaluate the structure of the interface,

- Low fidelity prototypes are better to discover more problems.
- Novice users are effective as expert users.
- Expert users are better to evaluate use flow
- Novice users are better to evaluate menu categorization
- Novice users are better to evaluate content

If the main aim is to measure performance data,

- High fidelity prototypes are better. Users are freer to tour between pages especially with fully interactive ones. Because the low fidelity prototypes do not provide complete connection between all points in the system and users could have some problems to focus on to completing task while waiting for the next screen load with paper prototype (Tam, 2006).
- Expert users can measure better. Because novices are unfamiliar with the system, they can raise the level of noise.

If there is a time restrictions,

- Low fidelity prototypes with either experts or novices can be used in usability tests. In some cases there is no significant difference between the results from both user groups. In addition low fidelity prototypes require less design effort, are less costly, easily editable and pratic to provide quick responses.

In usability literature, it has been mentioned in many studies that low fidelity prototypes are often accepted in the early design process. Because they are known as replicas of design ideas and with the sketchy appearance, they look like that they are unfinished. However, in this thesis, with the same approach as Virzi et al.(1996), the fidelity of the prototypes were aimed to be investigated at the later of design process with using the released product as a high fidelity one like studies by Lim et al.(2006) and Sauer et al.(2010) and paper version which was prepared from the original interface.

There is a difference in mentality between user groups. Novice users are observed to behave after comprehensive investigation and try to decide the required way and consider on understanding what they are doing while they are acting. On the other hand, the behavior logic of expert users is more action oriented and they already know what they are looking for. Traditionally, the tests conducted with novice users measure the learnability; with experts measure the optimal use.

Based on the findings of the current study, both user groups with different experience level and prototypes with different fidelity provide different results. The quality of results depends on how these factors combined with each other based on the research goal. Thus, first of all, it is important to clearly identify the purpose of the research.

5.2 Limitations of Study

There were some factors that cause limitations in this study should also be considered while evaluating the results. First of all, in this study, usability tests were evaluated and usability problems were categorized only one person who also moderated the tests. The data was gathered with only one perspective and the results could be subjective in some cases. In usability literature, it is suggested to use more evaluators to analyze the test sessions to minimize the risk of being biased. Beside this, two usability researchers rated the severity of problems (that is more subjective than problem categorization) because of the time constraints and resource limits. For this analyze, to increase the quality mean of the severity ratings, more evaluators were needed for many practical purposes. According to Nielsen, rating results from three evaluators is satisfactory (1995a).

The number of participants was also another issue in this study to be addressed as constraints in results. Due to the time and budget limitations, only five participants were used in each group. The sample size is sufficient to uncover 85% usability issues (Nielsen, 2000). In this study it was also observed that after some amount of users, adding another user to process provides only small amount of information. However there were some significant results in this study, it is possible that, if a large number of participants were attended to tests, the results might have different (there are no significant results with the severity of problems).

The other limitation about participants was with their language. It was mentioned in section 3.3 that almost half of tests were done in Germany with Turkish people currently living there. Five of these participants have been living in Germany for a long time and had confusion to understand some Turkish words. Therefore, this situation might influence on their performance.

The test environment in this research was also the source of some possible limitations. In this study, it was aimed to simulate the interaction mostly before to start driving. The tasks also include some duties that can be done while driving. As it was mentioned in section 3.5, usability tests could not be conducted neither in real environment nor in fully designed laboratory. In this case, the impact of environment context could not be included into the analysis.

In addition, almost all tests were done in weekdays and because of a lack of time and working hours of participants; they were visited in their own workplaces or homes. Thus, participants had to make some pauses during the test sessions for external factors. This might also influence on their concentration to complete the tasks.

The camera setup that was used in the sessions was mounted on the table (In section 3.5, Figure 3.6) to record the interaction between participants and prototypes. The behaviors of some participants might be biased due to the location of the camera.

In this study, due to the paper prototype, it was not found relevant results regarding the aesthetic part of the type “aesthetic and visual”. But, in another study with other variables, the results might be relevant.

5.3 Further Studies

The subject of this thesis; whether the fidelity of prototypes and user expertise effects on usability testing outputs especially on uncovered usability problems has no end to evaluate. There are several suggestions for future studies in this field. Sauer et al. (2010) worked with a physical product evaluated these two factors together. In this study, similar study was conducted on a digital interface.

To gather more data to make the findings widely acceptable and usable as a guide in usability tests, more experiments with both digital and physical products should be conducted with a larger amount of participants for significant results.

Furthermore, more evaluators should take part in the analysis at least two for categorical analysis of usability problems and at least three evaluators for rating the severity. In addition, it is suggested to use an extra observer to watch the sessions synchronously for more efficiency.

Another suggestion for further studies is to choose participants demographically equal to reduce the influence of differences. In addition, it was mentioned as a limitation that, if there is no possibility to conduct usability test in a lab, the test place were carefully chosen to reduce the external disturbances.

In this study, two main resources were used to reveal usability problems; observation and participants' comments. It is requested for further studies to compare also the results of these two sources.

REFERENCES

Beaudouin-Lafon, M., Mackay, W. (2002).Prototyping tools and techniques. In *The Human- computer interaction handbook.* pp 1006-1031 Hillsdale, NJ: Erlbaum.

Bruno, V., Al-Qaimari, G. (2004). Usability Attributes: An Initial Step Toward Effective User-Centred Development. *OZCHI,* Wollongong, Australia, November.

Dillon, A. and Song, M. (1997). An empirical comparison of the usability for novice and expert searchers of a textual and a graphic interface to an art-resource database. *Journal of Digital Information,* 1(1)

Dumas, J. S., Redish, J.C. (1999). A practical guide to usability testing. UK: Intellect Books Exeter (pp 22,26)

Eger, N., Ball, L. J., Stevens, R. and Dodd, J. (2007). Cueing retrospective verbal reports in usability testing through eye-movement replay. *In Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI… but not as we know it,* 1, Swinton, UK pp.129-137.

Faulkner, L. and Wick, D. (2005). Cross-user analysis: Benefits of skill level comparison in usability testing. *Interacting with Computers,* 17(6), 773-786

Genise, P. (2002). Usability Evaluation: Methods and Techniques. Date retrieved: 19.04.2015, address: http://en.wikipedia.org/wiki/Comparison_of_ usability_ evaluation_methods

Gerardo, J. L. S. (2007). The effectiveness of novice users in usability testing. (Master Thesis), Retrieved from: https://www.duo.uio.no/bitstream/ handle/10852/ 9681/1/Gerardo.pdf

Gray, M., Wardle, H. (2013). Observing gambling behaviour using think-aloud and video technology: methodological review, NatCen Social Research,

Hartson, H. R., Andre, T. S., Williges, R. C. (2001). Criteria for Evaluating Usability Evaluation Methods. *International Journal of Human-Computer Interaction* 13, 373-410

Hertzum, M., (2006). Problem Prioritization in Usability Evaluation: From Severity Assessments toward Impact on Design. *Internal Journal of Human-Computer Interaction,* 21 (2), 125-146

International Organization for Standardisation (1998). *Human Centered Design Process for Interactive Systems.* ISO 13407.

Ivory, M. Y. (2001). An Empirical Foundation for Automated Web Interface Evaluation. (Doctoral dissertation), Retrieved from http://webtango. berkeley.edu/papers/ thesis/thesis.pdf

Landis, J. R. and Koch, G. G. (1977) The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 13(1), 159-174

Lavery, D., Cockton, G., Atkinson, M. P. (1997). Comparison of evaluation methods using structured usability problem reports. *Behaviour & Information technology* 16 (4/5), 246-266, (p. 254)

Lim, Y. –K., Pangam, A., Periyasami, S., and Aneja, S. (2006).Comparative analysis of high- and low-fidelity prototypes for more valid usability evaluations of mobile devices. *In Proceedings of the 4th Nordic conference on Human- computer interaction: changing roles*. New York, NY, USA pp.291-300.

Lim, Y., Stolterman, E. and Tenenberg, J. (2008). The Anatomy of Prototypes: Prototypes as Filters, Prototypes as Manifestations of Design Ideas. *ACM Transactions on Computer-Human Interaction,* 15 (2), 7

Liu, Y., Osvalder, A.-L., and Karlsson, M. A. (2010). Considering the Importance of User Profiles in Interface Design, *User Interfaces, Rita Matrai (Ed.), ISBN: 978-953-307-084-1, InTech, DOI: 10.5772/8903*. Available from: http://www.intechopen.com/books/user-interfaces /considering-the-importance-of-user-profiles-in-interface-design

Lundberg, J. (2010). Guidelines for Developing an Interactive Multimedia Prototype: Based on comparison of Low-and-High-fidelity prototypes in usability testing. (Master Thesis), Retrieved from: https://www. nada.kth.se/utbildning/grukth/ exjobb/rapportlistor/2010/ rapporter10/ lundberg_johan_10005.pdf

Macnamara, J. (2005). Media content analysis: Its uses, benefits and best practice methodology. *Asia Pasific Public Relations Journal,* 6(1), 1-34

Magnussen, J. C. (2010). Prototypes in usability testing: the implications o richness in interaction fidelity. (Master Thesis), Retrieved from: https://www.duo.uio.no/bitstream/handle/10852/8746/Magnussen.pdf?sequence=4&isAllowed=y

Mäuselein, M. (2007). Paper Prototypes vs. Computer-based Prototypes in a User-centered Design Process. (Master Thesis), Retrieved from: http://www.cs.uni-paderborn.de/fileadmin/Informatik/FG-Szwillus/ Diplom-Masterarbeiten/MaeuseleinDA.pdf

McCurdy, M., Connors, C., Pyrzak, G., Kanefsky, B. and Vera, A. (2006). Breaking the barrier: an examination of our current characterization of prototypes and a example of mixed-fidelity success. In *Proceedings of the SIGCHI Conference on Human factors in Computing Systems (CHI' 06).* ACM Press, New York, NY, 1233-1242

Nielsen, J. (1993). Usability Engineering. Morgan Kaufmann, San Francisco, CA.(p 177)

Nielsen, J. (1995a). Severity Ratings for Usability Problems. Date retrieved: 19.04.2015, address: http://www.nngroup.com/articles/how-to-rate-the-severity-of-usability-problems/

Nielsen, J. (1995b). Summary of Usability Inspection Methods. Date retrieved: 19.04.2015, address: http://www.nngroup.com/articles/summary-of-usability-inspection-methods/

Nielsen, J. (1995c). 10 Usability Heuristics for User Interface Design. Date retrieved: 19.04.2015, address: http://www.nngroup.com/articles/ten-usability-heuristics/

Nielsen, J. (2000). Why You Only Need to Test with 5 Users. Date retrieved: 19.04.2015, address: http://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/

Nielsen, J. (2001). Success rate: The Simplest Usability Metric. Date retrieved: 19.04.2015, address: http://www.nngroup.com/articles/success-rate-the-simplest-usability-metric/

Nielsen, J. (2011). Parallel&Iterative Design + Competitive Testng = High Usability. Date retrieved: 19.04.2015, address: http://www.nngroup.com/articles/parallel-and-iterative-design/

Nielsen, J. (2012a). Usability 101: Introduction to Usability. Date retrieved: 19.04.2015, address: http://www.nngroup.com/articles/usability-101-introduction-to-usability/

Nielsen, J. (2012b). How Many Test Users in a Usability Study. Date retrieved: 19.04.2015, address: http://www.nngroup.com/articles/how-many-test-users/

Nielsen Norman Group (2014). Turn User Goals into Task Scenarios for Usability Testing. Date retrieved: 19.04.2015, address: http://www.nngroup.com/articles/task-scenarios-usability-testing/

Partala, T., Kangaskorte, R., (2009). The Combined Walkthrough: Measuring Behavioral, Affective, and Cognitive Information in Usability Testing. *Journal of Usability Studies* 5(1), 21-33

Petrie, J. N., Schneider, K. A. (2007). Mixed –Fidelity Prototyping of user Interfaces. *Lecture Notes in Computer Science,* 4323, 199-212

Preece, J., Roggers, Y. and Sharp, H. (2002). Interaction design: beyond human-computer interaction, New York (p 169, 245)

Rettig, M. (1994). Prototyping for tiny fingers. *Communications of the ACM,* 37(4), 21-27

Sauer, J., Franke, H., Ruettiner, B. (2008). Designing interactive consumer products: Utility of paper prototypes and effectiveness of enhanced control labeling. *Applied Ergonomics,* 39, 71-85

Sauer, J., Sondereger, A. (2009). The influence of prototype fidelity and aesthetics of design in usability test: Effects on user behaviour, subjective evaluation and emotion. *Applied Ergonomics,* 40(4), 670-677

Sauer, J., Seibel, K., Ruettiner, B. (2010). The influence of expertise and prototype fidelity in usability tests. *Applied Ergonomics,* 41, 130-140

Sauro, J. (2011). How to find the right sample size for a usability test. Date retrieved: 19.04.2015, address: http://www.measuringu.com/blog/sample-size-problems. php

Sauro, J. (2013). Rating the Severity of Usability Problems. Date retrieved: 19.04.2015, adress: http://www.measuringu.com/blog/rating-severity.php

Sauro, J. (2014). The Relationship Between Problem Frequency and Problem Severity in Usability Evaluations. *Journal of Usability Studies,* 10 (1), 17-25

Sefelin, R., Tscheligi, M., Giller, V. (2003). Paper Prototyping- What is it good for?: A comparison of paper-and computer-based low-fidelity prototyping. In *Proceedings of the Extended Abstracts on Human Factors in Computer Systems (CHI' 03).* Ft. Lauderdale, FL. ACM Press, New York, NY, 778-779

Schade, A. (2015). Pilot Testing: Getting It Right (Before9 the First Time. Date retrieved: 19.04.2015, adress: http://www.nngroup.com/articles/pilot-testing/

Shluzas, L. A., Sadler, J., Currano, R. M., Sanks, T., Steinert, M. and Katila, R. (2013). Comparing Novice and Expert User Inputs in Early Stage Product Design. *IASDR 2013 (5th International Congress of International Association of Societies of Design Research)* . *Design Society; Tokyo*

Snyder, C. (2003). Paper prototyping: The fast and easy way to design and refine user interfaces. Morgan Kaufmann [Pdf slides]. Retrieved from http://www2.engr.arizona.edu/~ece596c/lysecky/uploads/Main/Lec6.pdf

Sonderegger, A. (2010). Influencing Factors in Usability Tests: The testing Situation, the Product Prototype, and the Test User. (Doctoral dissertation), Retrieved from http://doc.rero.ch/record/28385/files/SondereggerA.pdf

Students in the Master of technical and Scientific Communication Program. (2004). *Usability Testing: Developing Useful and usable Products*. Miami University of Ohio

Tam, M. (2006). Using Paper Prototyping as a usability testing methodology for web application development. (Master Thesis), Retrieved from: http://summit.sfu.ca/system/files/iritems1/6085/etd2568.pdf

Travis, D. (2014). 247 web usability guidelines. Date retrieved: 19.04.2015, adress: http://www.userfocus.co.uk/resources/guidelines.html

Tullis, T., Albert, W. (2008). Measuring the User Experience: Collecting, Analyzing and Presenting Usability Metrics. Morgan Kaufmann [Pdf slides]. Retrieved from http://www2.engr.arizona.edu/~ece596c/lysecky/uploads/Main/Lec11.pdf

Umar, A., Tatari, K. K. (2008). Appropriate Web Usability Evaluation Method during Product development: A comparison and analysis of formative web usability evaluation methods (Master Thesis), Retrieved from

http://www.bth.se/fou/cuppsats.nsf/all/0ba947e15907c31cc125741100517192/$file/MSE_2008_03_Final_Update.pdf (Thesis no: MSE-2008-03)

Virzi, R. A., Jeffery, L. S., Karis, D. (1996). Usability problem identification using both low-high-fidelity prototypes. In *Proceedings of the SIGCHI Conference on Human factors in Computing Systems (CHI' 96).* ACM Press, New York, NY, 236-243

Walker, M., Takayama, L., and Landay, A. (2002).High-Fidelity or Low Fidelity, Paper or Computer? Choosing Attributes When Testing Web Prototypes. *In Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting*. Baltimore, USA, HFES, Santa Monica pp. 661–665.

Ziefle, M. (2002). The influence of user expertise and phone complexity on performance, ease of use and learnability of different mobile phones. *Behaviour & Information Technology*, 21(5), 303-311

URL-1< http://www.usabilityfirst.com/about-usability/introduction-to-user-centered-design

URL-2 < http://www.usabilitybok.org/summative-usability-testing

URL-3 < http://research.cs.vt.edu/usability/projects/uaf% 20and% 20tools/upc.htm > date retrieved 29.03.2015

URL-4<http://www.boardofinnovation.com/wp-content/uploads/2013/04/paper protoyping. jpg

APPENDICES

APPENDIX A: Usability testing scenarios

APPENDIX B: Screen Samples of Low and High fidelity prototypes

APPENDIX C: Statistics

APPENDIX D: Usability problem severity ratings guideline

APPENDIX A

Usability testing scenarios- English

Task 1- Updating home address

a) Navigate to your home which is already saved in the device.

b) Change your current address with the new one written on the paper that was given to you..

" Ringstrasse 2, Grossmehring"

Task 2- Adding favorite address

Imagine that you are travelling a lot for your work. You don't want to enter all these most travelled addresses again and again. This device has a function that you can save your addresses with a special name as a favorite, such as; mom's home, my doc. etc.

a) Save your work address with a name "iş yerim (my work)" (this address is in the recent destination list)

b) Read the message about the task completeness.

Task 3- Adding a sub-route (POI) into current route

a) Navigate to your work address that you have just added in the list as a favorite with a name "iş yerim"

Imagine that your gas is running out and need to fill your car immediately. You can update your current route to visit the selected gas station not to change your final destination.

b) Add a nearest gas station into your current route "iş yerim"

Task 4 – Creating Quick menu

You are driving back to your home from work and it is dark outside.

a) Change your screen mood into night colors.

Imagine that you make this setting every day. This device has a function that you can collect some functions (frequently used) into a special menu and you can easily reach this menu while you are on the map screen.

b) Create this menu for quick settings and add "day/night colors" and "sound on/off" functions.

Task 5 – Explaining all the information on the main screen (map) and changing a desired setting

Imagine that you are driving again in daytime.

a) Change your screen mood into day colors. Find this function into the quick menu that you have already created.

b) Explain all the information on the main screen(map)

c) Change the time and distance units (“18:00” to “6pm” / “km” to “mil”)

Task 6 – Warn setting for speed limit

This device gives some information about your destination on the map screen. Beside this, you are also informed with a sound in some situations, such as; exceed the speed limit.

a) Set the audio warning to be warned if you drive faster than allowed in current region.

Kullanıcı testi senaryoları- Türkçe

İşlem 1- Ev adresini güncellme

a) Sistemde kayıtlı olan ev adresinize gitmek üzere rotanızı çiziniz.

b) Mevcut ev adresinizi size verdiğimiz kağıtta yazan yeni adres ile değştiriniz.

" Ringstrasse 2, Grossmehring"

İşlem 2- Favori adres ekleme

İşiniz gereği sürekli seyahat ediyorsunuz. Her defasında aynı adresleri tekrar tekrar girmek yerine bu aletinn içinde bunları özel isimler vererek kaydedebiliyorsunuz. Örneğin annemin evi, doktorum..vs.

a) İş yeri adresinizi "iş yerim" olarak kaydediniz. (bu adres daha once gittiğiniz adresler arasında mevcut)

b) İşlemi gerçekleştirebildiğiniz ile ilgili çıkan bilgi mesajını okuyunuz.

İşlem 3- Mevcut rotaya ara adres eklemek

a) "iş yerim" olarak az once eklediğniz adresi ilgili bölümden bularak iş yerinize gitmek üzere rotanızı çiziniz

Yola çıktınız ve benzininizin bitmek üzere olduğunu farkettiniz.

b) İş yerinize gitmek üzere çizdğiniz rotayı iptal etmeden size en yakın benzinliği mevcut rotanıza ekleyerek rotanızı güncelleyebilyorsunuz. Bu işlemi gerçekleştiriniz.

İşlem 4 – Kısa yol menüsü oluşturma

İş yerinizden evinize gitmek üzere yola çıktığınızı ve havanın karardığını düşünün.

a) Cihazınızın kullanım modunu gece kullanımı için değiştiriniz.

Bu ayarı her gün yaptığınızı düşünün. Cihazınız içinde bunun gibi sık kullandığınız bazı ayarları bir araya toplayabileceğiniz bir menu yaratabiliyorsunuz ve harita ekranındayken bu menüye hızlıca ulaşabiliyorsunuz.

b) Bu işlem için gerekli menüyü oluşturunuz. Bunun içine "Ses ac/kapa" ve "gece-gündüz renk" değişimi fonksiyonlarını ekleyiniz.

İşlem 5 – Harita ekranı bilgi okuma-güncelleme

Tekrar gündüz yolculuk yaptığınızı düşünün.

a) Cihazınızın kullanım modunu tekrar gündüz kullanımı için değiştiriniz. Bu değişikliği yapmak için ilgili fonksiyonu biraz önce oluşturduğunuz "hızlı ulaşım" menüsü içinden açınız.

b) Harita ekranındaki bilgileri bulup sesli bir şekilde söyleyiniz.

c) Saat ve mesafe birim değişikliğini yapınız.("18:00"- "06pm" / "km"-"mil")

İşlem 6 – Hız limiti sesli uyarı ayarı

Navigasyon cihazınız sürüşünüzle ilgili bilgileri ekranda göstermenin yanı sıra özellikle hız limitini aştığınız gibi bir durumda sizi sesli olarak da uyarma özelliğine sahiptir.

a) İzin verilenden daha hızlı sürdüğünüzde cihazınızın sizi sesli olarak uyarması için ilgili ayarını yapınız.

APPENDIX B

Screen Samples of Low and High fidelity prototypes

Figure B.1: Map screen

Figure B.2: Main menu

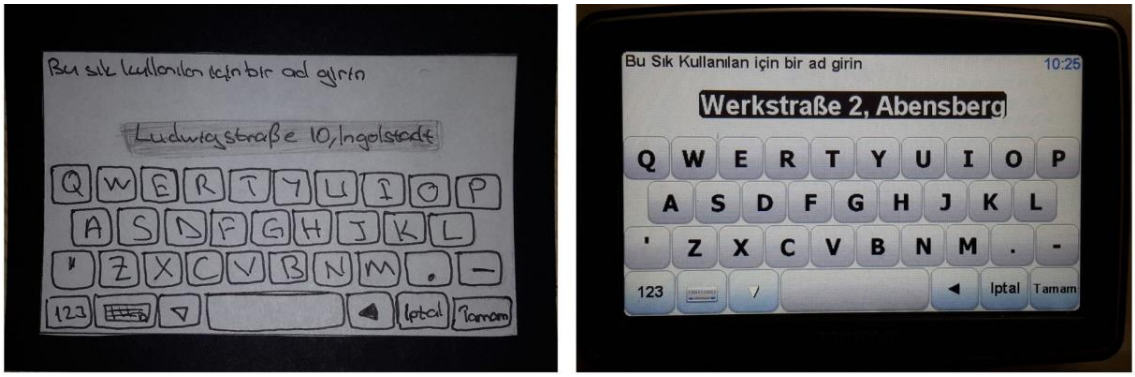

Figure B.3: Updating name of the address

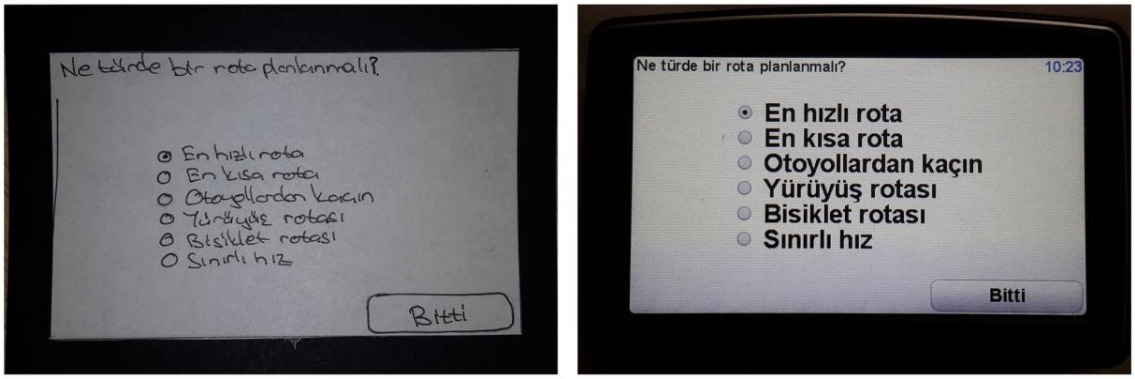

Figure B.4: Route options

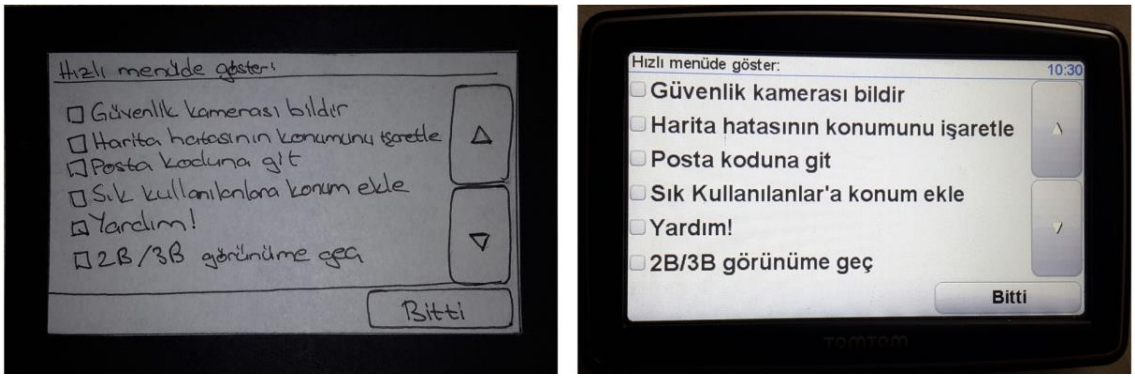

Figure B.5: Quick menu list

Figure B.6: Menu with inactive items

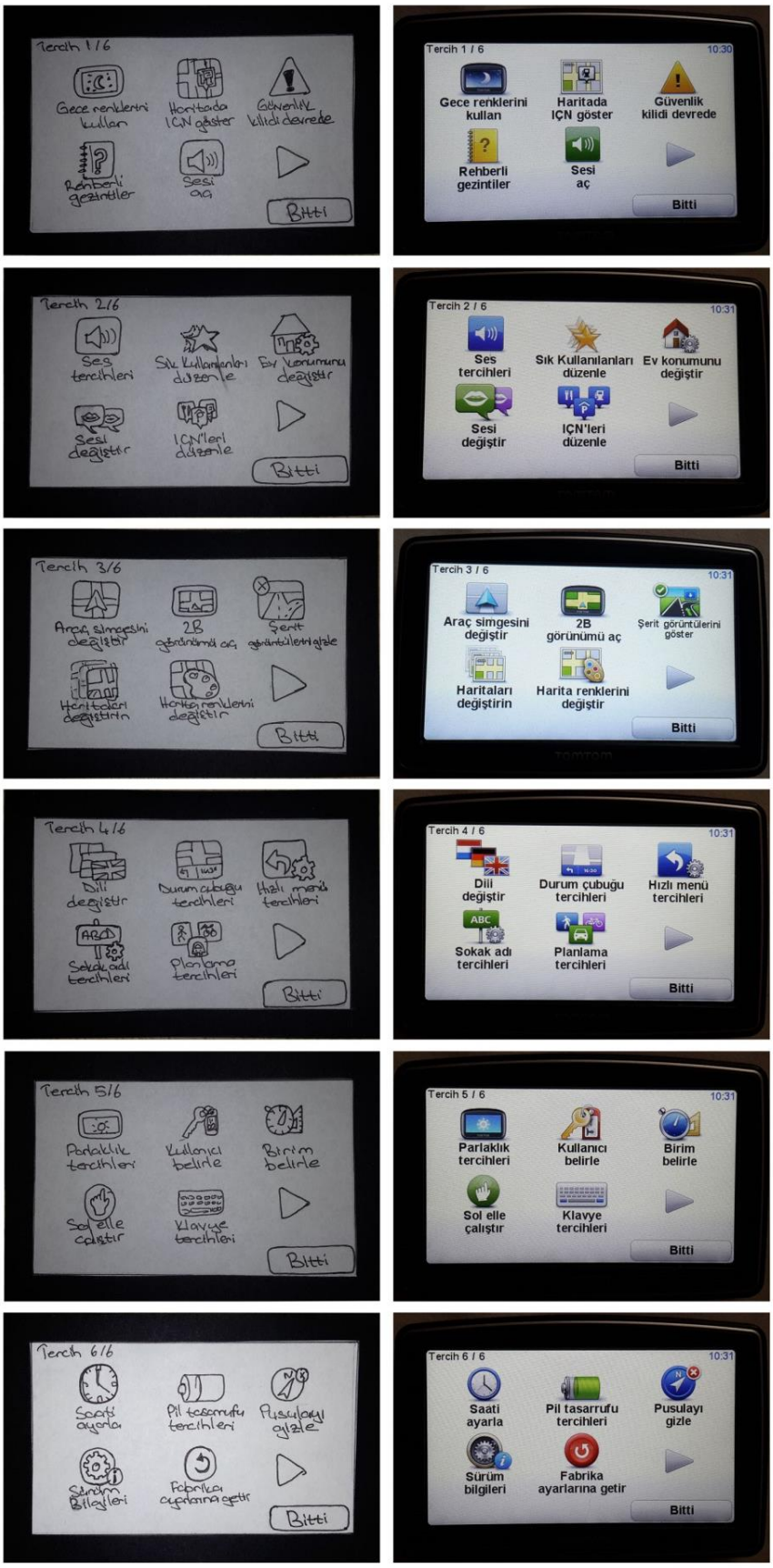

Figure B.7: Preferences menu

APPENDIX C

Statistics

Table C.1: Success rate

Tests of Between-Subjects Effects

Dependent Variable: successrate

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	11,800[a]	3	3,933	,894	,466
Intercept	1729,800	1	1729,800	393,136	,000
Fidelity	1,800	1	1,800	,409	,531
Expertise	9,800	1	9,800	2,227	,155
Fidelity * Expertise	,200	1	,200	,045	,834
Error	70,400	16	4,400		
Total	1812,000	20			
Corrected Total	82,200	19			

Table C.2: Time on task

Tests of Between-Subjects Effects

Dependent Variable: Timeontask

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	241034,800[a]	3	80344,933	2,098	,141	,282
Intercept	12736080,000	1	12736080,000	332,614	,000	,954
Fidelity	16245,000	1	16245,000	,424	,524	,026
Expertise	217569,800	1	217569,800	5,682	,030	,262
Fidelity * Expertise	7220,000	1	7220,000	,189	,670	,012
Error	612653,200	16	38290,825			
Total	13589768,000	20				
Corrected Total	853688,000	19				

a. R Squared = ,282 (Adjusted R Squared = ,148)

Table C.3: Number of Problems

Tests of Between-Subjects Effects

Dependent Variable: Number of Usability problems

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	115,750[a]	3	38,583	2,180	,130	,290
Intercept	11092,050	1	11092,050	626,669	,000	,975
Fidelity	8,450	1	8,450	,477	,500	,029
Expertise	101,250	1	101,250	5,720	,029	,263
Fidelity * Expertise	6,050	1	6,050	,342	,567	,021
Error	283,200	16	17,700			
Total	11491,000	20				
Corrected Total	398,950	19				

a. R Squared = ,290 (Adjusted R Squared = ,157)

Table C.4: Severity of problems

Tests of Between-Subjects Effects

Dependent Variable: Severity rate

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	,074[a]	3	,025	1,495	,254
Intercept	113,621	1	113,621	6901,834	,000
Fidelity	,038	1	,038	2,299	,149
Expertise	,031	1	,031	1,896	,188
Fidelity * Expertise	,005	1	,005	,292	,596
Error	,263	16	,016		
Total	113,959	20			
Corrected Total	,337	19			

Table C.5: Types of problems

Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
	Content	39,200	1	39,200	4,519	,049	,220
	Use flow	1,800	1	1,800	1,075	,315	,063
	Page Layout	,800	1	,800	1,231	,284	,071
	Menu categorization	,800	1	,800	,762	,396	,045
Fidelity	Interactive Components	1,250	1	1,250	,510	,485	,031
	System status and Response	,200	1	,200	,500	,490	,030
	Aesthetic and visual	9,800	1	9,800	17,043	,001	,516
	Content	12,800	1	12,800	1,476	,242	,084
	Use flow	12,800	1	12,800	7,642	,014	,323
	Page Layout	,200	1	,200	,308	,587	,019
	Menu categorization	5,000	1	5,000	4,762	,044	,229
Expertise	Interactive Components	26,450	1	26,450	10,796	,005	,403
	System status and Response	,000	1	,000	,000	1,000	,000
	Aesthetic and visual	5,000	1	5,000	8,696	,009	,352
	Content	20,000	1	20,000	2,305	,148	,126
	Use flow	1,800	1	1,800	1,075	,315	,063
	Page Layout	,800	1	,800	1,231	,284	,071
	Menu categorization	3,200	1	3,200	3,048	,100	,160
Fidelity * Expertise	Interactive Components	,050	1	,050	,020	,888	,001
	System status and Response	,200	1	,200	,500	,490	,030
	Aesthetic and visual	3,200	1	3,200	5,565	,031	,258
	Content	138,800	16	8,675			
	Use flow	26,800	16	1,675			
	Page Layout	10,400	16	,650			
	Menu categorization	16,800	16	1,050			
Error	Interactive Components	39,200	16	2,450			
	System status and Response	6,400	16	,400			
	Aesthetic and visual	9,200	16	,575			

APPENDIX D

Usability problem severity ratings guideline

High severe problems

- prevent the task completion
- must be fixed before product can be released.
- cause extreme irritation on user

Medium severe problems

- limit the task completion
- important be fixed before product can be released.
- cause moderate irritation on user

Low severe problems

- cause minor effects on task completion
- can be fixed if there is enough time before product can be released.
- cause minimum irritation on user

CURRICULUM VITAE

Name Surname:	Gamze KAYA KAPLAN
Place and Date of Birth:	Yeşilova – Turkey / 09.06.1987
Address:	Ringstrasse 2, 85098 Grossmehring / Germany
E-Mail:	gamzekya@gmail.com
B.Sc.:	Istanbul Technical University
M.Sc.:	Istanbul Technical University

Professional Experience and Rewards:

List of Publications and Patents: