Aalto University
School of Science
Department of Computer Science and Engineering

A/B testing and usability assessment methods in small companies

Master’s Thesis

Espoo

10.5.2014

Maksim Luzik

Supervisor: Marjo Kauppinen

Instructor: Marjo Kauppinen
Charlotta Turku

AALTO UNIVERSITY SCHOOL OF SCIENCE Department of Computer Science and Engineering	ABSTRACT OF THE MASTER'S THESIS

Author: **Name of the thesis:**	Maksim Luzik A/B testing and usability assessment methods in small companies	
Date:	10. May 2014	**Number of pages:** 57
Major:	User interfaces and usability	**Code:** T242-3
Supervisor: **Instructor:**	Prof. Marjo Kauppinen Prof. Marjo Kauppinen MSc Charlotta Turku	

Usability assessment methods have been considered for long as useful tools to improve usability of the product or service. Many small companies are still reluctant to practice usability assessment methods due to many reasons including limited resources and management bias. A/B testing is a new movement in a corporate world. Multiple companies are considering to begin practicing the method or are practicing it already. Moreover, some people assume that A/B testing can actually cover all the usability assessment methods.

This work provides answers to what A/B testing and usability assessment methods are and how they can be practiced in a small company (10-100 employees). Additionally the strengths and weaknesses of both methods are presented and compared. Finally suitability of the methods will be discussed.

The target of optimization in this paper is an application of A/B test into a complex enterprise system. Only common usability assessment methods will be evaluated that are used today. These include standard usability testing, heuristic evaluation and cognitive walkthrough. No other usability assessment methods are included in this paper.

Both practices: usability assessment methods and A/B testing have a place in a project. Despite their different approach they often complement each other and give answers to different questions. Usability assessment methods tend to answer to question *why* while A/B test usually answers to question *which* or *how many*. Usability assessment methods work well as an initial evaluation while A/B test can be used as a follow up method to fine-tune product or service even further.

Keywords:	A/B test, split test, usability assessment, usability inspection, usability testing, heuristic evaluation, cognitive walkthrough, conversion, conversion rate optimization, Kiosked
Language:	English

AALTO-YLIOPISTO
TEKNILLINEN KORKEAKOULU
Informaatio- ja luonnontieteiden tiedekunta
Tietotekniikan tutkinto-ohjelma

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä:	Maksim Luzik	
Työn nimi:	A/B-testaus ja käytettävyyden arviointimenetelmät pienissä yrityksissä	
Päiväys:	10.05.2014	**Sivumäärä:** 57
Pääaine:	Käyttöliittymät ja käytettävyys	**Koodi:** T242-3
Valvoja:	Prof. Marjo Kauppinen	
Ohjaaja:	Prof. Marjo Kauppinen DI Charlotta Turku	

Käytettävyyden arviointimenetelmät ovat olleet pitkään hyödyllisiä työkaluja yritysten tuotteiden jalostuksessa. Vaikka osa yrityksistä käyttääkin käytettävyyden arviointimenetelmiä omassa liiketoiminnassaan, monet pienet yritykset ovat edelleen haluttomia harjoittamaan menetelmiä eri syistä. A/B-testaus on uusi villitys yritysmaailmassa. Monet yritykset harkitsevat A/B testin käyttöönottoa tai ovat jo siirtyneet käyttämään sitä. Jotkut jopa olettavat, että A/B testi korvaa perinteiset käytettävyyden arviontimenetelmät.

Tämä työ käsittelee yleisesti A/B-testausta ja käytettävyyden arviointimenetelmiä sekä sitä, miten ne eroavat toisistaan. Työssä sivutaan myös menetelmien käytännön toteutusta pienissä yrityksissä (10-100 työntekijää). Lisäksi A/B-testauksen ja käytettävyyden arviointimenetelmien vahvuudet ja heikkoudet tuodaan esille. Lopuksi arvioidaan käytettävyyden arviointimenetelmien ja A/B-testauksen käyttötilannetta.

Tämä työ keskittyy A/B-testauksen soveltamiseen yrityksen monimutkaiseen järjestelmään. Vain yleiset käytettävyyden arviointimenetelmät, joita ovat käytettävyystesti, heuristinen arviointi ja kognitiivinen läpikäynti, käydään läpi tässä julkaisussa. Mitään muita käytettävyyden arviointimenetelmiä ei käsitellä tässä työssä.

Molemmat menettelyt, käytettävyyden arviointimenetelmät ja A/B-testaus, ovat arvokkaita työkaluja projekteissa. Huolimatta siitä, että menetelmien lähestymistavat eroavat toisistaan, ne usein täydentävät toisiaan ja antavat vastaukset eri kysymyksiin. Käytettävyyden arviointimenetelmät yleensä antavat vastauksen kysymykseen *miksi*, kun puolestaan A/B-testaus pyrkii vastaamaan kysymykseen *kumpi* tai *kuinka monta*. Käytettävyyden arviointimenetelmät toimivat hyvin ensimmäisenä arviointivaiheena, kun taas A/B-testaus sopii hyvin tuotteen tai palvelun hienosäätöön.

Avainsanat:	A/B-testaus, A/B-testi, käytettävyyden arviointimenetelmät, käytettävyystesti, heuristinen arviointi, kognitiivinen läpikäynti, konversio, konversioasteen optimointi, Kiosked
Kieli:	englanti

Acknowledgements

I would like to thank my supervisor and instructor Marjo Kauppinen for supporting and assisting me with this paper despite the fact that it took almost two years to finish it. Also I would like to give my gratitude to my spouse and family for pushing me to write and finalize this thesis.

Espoo 10.5.2014 Maksim Luzik

List of Abbreviations

API
Application programming interface; specifies how different applications should communicate with each other. It defines specific rules and dictates the standard approach for communication.

CEO
Chief executive officer; Is the highest ranking person in organization who reports to the board of directors.

CRO
Conversion rate optimization; An internet marketing practice to improve the conversion rate of website visitors to customers.

LPO
Landing page optimization; LPO is a practice for improving the web page that user lands on first when following referenced link. In LPO idea is to convert a visitor to customer. One of the methods used to achieve this is usually A/B testing.

SaaS
Software as a Service; a software delivery model where software and the data are hosted centrally in the cloud. Sometimes also known as software “On-Demand”.

SEO
Search Engine Optimization; a practice where website or product is optimized for search engines so they can more easily find the website and rank it higher in their search results.

UE
User experience; User experience involves user perceptions and emotions when using particular product, system or service.

UI
User interface; an element in between of the human-computer-interaction. The layer that user is using to control the system.

UX
User experience; see abbreviation UE

WYSIWYG
What you see is what you get; A user interface or system where content is displayed as is. User can see the final view of the content while editing it simultaneously.

Table of Contents

1 Introduction

Everyone knows that good usability and conversion in web from a visitor to a client is critical for a company to grow. Regrettably some companies consider that usability and conversion increase can happen by itself without putting extra effort in those areas. Frequently the managers think that they are the experts in doing all the decisions regarding usability or conversion optimization. Often the reason behind absence of proper usability and conversion improvements is the lack of information and knowledge about methods in achieving better user experience. In many cases companies concentrate on conversion optimization first and begin to pay attention to usability improvements in very late stages.

The purpose of this thesis is to describe what usability and conversion optimization is. Furthermore most commonly used methods for improving usability and conducting conversion optimization will be assessed. Also strengths and weaknesses of both approaches will be discussed. In empirical part one qualitative and three quantitative research results will be presented related to usability and A/B testing. Finally usability assessment methods and A/B testing method will be analyzed and discussed.

Focus on small companies was chosen due to lack of well-established processes in those companies. Many small companies do not have a proper usability team nor dedicated resources to run extensive conversion optimizations. The goal of this paper is to help companies and people to gain smoother access to world of usability and conversion.

2 Research problems and questions

The objective of this thesis is to provide information for the companies about why both usability assessment methods and A/B testing are important tools to enhance the usability of the product or service. The main research focus of the topic is to examine how well A/B testing can be implemented and taken into use in a small company as well to study how A/B testing compare to the traditional usability assessment methods. Main issues are as follows:

1. What are the strengths and weaknesses of the usability assessment methods and A/B testing respectively?
2. How A/B tests are actually realized in small companies?

The sub-goal is to understand how A/B testing can supplement the usability testing for a small company. Grasping the strengths and weaknesses of the methods can help us to understand and recognize also situations where the methods should be used. Furthermore, explaining how a small company could conduct A/B tests and what should be taken into consideration when executing an A/B test.

3 Theoretical background

To understand how usability and A/B tests are connected, it is necessary to familiarize with usability assessment methods as well as with A/B testing practice. In addition it is essential to distinguish between two terms that are used today interchangeably: usability and user experience. In usability assessment methods, only usability testing and the main usability inspection methods will be covered to encompass basic principles of the methods.

One can't deny the connection between usability assessment methods and A/B testing. Even though mentioned methods are used in different situation and context they still overlap each other at some level. A/B testing can sound and feel more tempting and easier to approach than usability testing. In turn it can be also much more challenging.

3.1 Usability and user experience

The term usability is widely used and raised in many discussions. It is admirable that the concept has spread so widely from the early days of 1980s (Dumas & Salzman, 2006). The disturbing aspect occurs with misconception of what usability is and how it should be practiced.

ISO 9241–11 standard defines usability as "extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in specified context of use". Effectiveness denotes "accuracy and completeness with which user achieves specified goals". Efficiency is defined as "resources expended in relation to accuracy and completeness with which user achieve goals". Lastly satisfaction is signified as "freedom from discomfort, and positive attitudes towards the use of the product". (International Organization for Standardization, 1998)

Usability definition is thought by usability professionals to be too strictly limited and thus is sought to be improved (Dumas & Salzman, 2006). This is where user experience comes into play, to expand the understanding of usability concept and actual user experience. Hancock et al. (2005) approached customer requirements as Maslow's hierarchy of needs. In general functionality is essential, usability is important and lately also expected, and pleasurable interactions are what users actually want (Dumas & Salzman, 2006).

Functionality is an essential part of a system. Without functionality user can not complete the desired goals. Usability is an attribute that makes a system or product easy to understand and use. User experience is combination of user's aesthetics and individual, social and contextual factors (Dumas & Salzman, 2006).

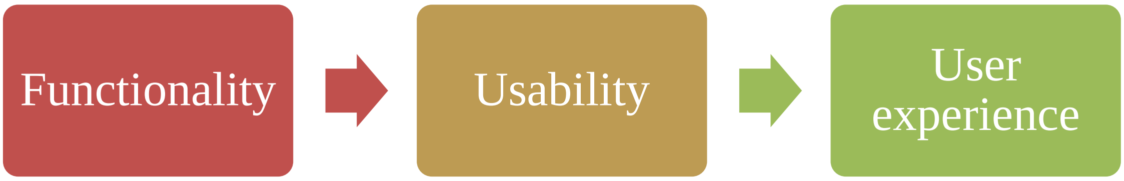


Figure 1: The relationship between functionality, usability and user experience

Without functionality, there can not be usability, without usability there can not be good user experience. Looking at Figure 1, one could see that bad user experience is possible if the usability is implemented poorly or the part is left out from the equation completely. To understand the concept better, consider following classical User Experience (UE) example setup (Nieminen, 2013):

- Functionality as roads that lead to desired destination
- Usability as an attribute that measures how easy it is for a driver to read the road signs.
- User experience as a road trip

A good user experience would require that there is a road that would lead driver to its destination (functionality); the road signs would indicate easy instructions on how to get to the desired location; there are right amount of the signs and if strayed the driver can easily find the way back (usability). One thing that usability specialist should remember is that some things, like user's aesthetics affect user experience drastically and are almost impossible to change, at least in a short period of time. Combining all of these previously mentioned attributes, we get user experience.

It is not false to assume that every startup company has struggled to get their first user base and increase their customer numbers even further. Getting new customers and improving your conversion rate from visitor to customer is never an easy task – with usability improvements it is possible to gain more positive user experience and thus more customers. Usability can be a key factor to differentiate your business from the competition. Especially in particular market forms, like monopolistic competition and oligopolies, where competition can be highly price driven. Usability can be the one and only difference between someone making a purchase from you or buying the product from the competitor.

3.2 Usability assessment methods

The most widely used methods to assess usability of a product or service are heuristic evaluation and traditional usability testing (Hollingsed & Novick, 2007). Even though these two methods already include interviews and perhaps some kind of questionnaires as well, the interview alone or well-planned questionnaire can provide useful information if used at the right time. Interviews and questionnaires are especially practical in user-centered product development.

The assessment methods were selected based on the citation frequency in usability assessment and evaluation related papers and articles. The selected methods are widely used to this day. In addition the amount of assessment methods were kept low in this paper and only the articles relative to this work were chosen. It was considered to include interviews and questionnaires into usability assessment methods as well, but they were dropped due to little importance in this paper. Often interviews and questionnaires are more tempting to use for usability evaluation or satisfaction level, but tend to require more discipline, commitment and work than usability inspection methods. The predefined practices in usability inspection methods often provide boundaries and tools to create a necessary assessment and analyze the feedback efficiently.

3.2.1 Usability inspection methods

In the early 1990 usability became a focus in product development (Hollingsed & Novick, 2007). With the focus on improving user experience many usability methods unfolded and became a de-facto standards for usability testing and evaluation of user experience. For instance, heuristic evaluation is a widely known informal usability method defined by Jakob Nielsen in 1990 and updated in 1994. It is still frequently used method in many companies and case studies (Hollingsed & Novick, 2007). Jakob Nielsen is a known name to every usability specialist or a person working in the field of interaction design and usability.

Some of the usability inspection methods have been introduced almost twenty years ago. Since then their use and effectiveness have been studied and put into practice. Some of the methods reached their dead ends, while others evolved or proved to be useful to this day. Heuristic evaluation and cognitive walkthrough remain still widely used techniques (Hollingsed & Novick, 2007). There are as well other popular methods used in user-centered design but they will not be covered in the following chapters. The usability inspection methods listed in this chapter were picked by their popularity and active state in companies and usability research field.

3.2.1.1 Cognitive walkthrough

Cognitive walkthrough is one of the popular usability inspection methods and is derived from the programming technique called code walkthrough[1] (Nielsen, 1994). In cognitive walkthrough design of the user interface is evaluated in exploratory approach, where interface is shown to group of users. Like many other inspection methods, cognitive walkthrough can be performed at any stage of a development process, from the original mock-ups through final release. The evaluation is done in two steps: a preparatory phase and an analysis phase. In the preparatory phase the evaluators select an interface to be used, a task to be completed, actions to be taken, and likely users that will be using the product (Hollingsed & Novick, 2007). In the analysis step, the experimenters use the following four steps to evaluate user interface (Polson, et al., 1991) :

- The goal to be completed within the system is set
- Currently available actions are determined
- The action is selected that is thought to take user closer to their goal
- The chosen action is performed and evaluated based on the feedback given by the system

As pointed out by (Wharton, et al., 1994, pp. 105-140) the old cognitive approach had two main limitations: the repetitiveness of filling out the forms and the limited problems the process found. Later a new approach in cognitive walkthrough addressed these limitations by evaluating interface in small groups instead of individual evaluators. Despite the improvement to the process the cognitive walkthrough drew criticism about unclear guidelines (Hollingsed & Novick, 2007). When using mock-ups or non-finished releases it is unclear for evaluators or in this case users to understand what kind of actions user can take to complete the goal. The problems occur in situations where user decides to take an action that has not been considered by the designer. When the non-existent action is chosen by user, the designer does not have a mock-up to show. After impasse the user needs to continue from a next pre-selected mockup, even though he/she did not choose that action. Cognitive walkthrough is particularly suited to evaluate designs before testing with users becomes feasible (Hertzum & Jacobsen, 2001).

Studies show that cognitive walkthrough can be accomplished by evaluators with experience in the process (Hollingsed & Novick, 2007). Secondly it is found out that the method is learnable and

[1] Code walkthrough is a technique where one coder shows and explains his code and decisions to a fellow coder. Together it is much more easier to find out problems, security holes or logical issues in the code than trying to review your own code.

usable for the novices as well. There are definitely own challenges when a novice is conducting cognitive walkthrough for the first time. The earlier the stage of the interface, the more difficult it is to arrange the walkthrough and thus requires the knowledge and experience of the designer to react to unexpected situations and keep the process on track. The usability specialist leading the cognitive walkthrough needs to avoid design discussions that are not relevant for the walkthrough tasks and goals as well. The scenario of the evaluation should be adequately described otherwise the evaluation could be non-effective (Hollingsed & Novick, 2007).

3.2.1.2 Heuristic evaluation

Heuristic evaluation is a method of usability analysis, where usability experts try to find issues and usability problems in a product or a system (Nielsen, 1990). Evaluation is done by looking at the interface and trying to come up with opinions about what is good and what is bad. Most user interface evaluations are actually heuristic evaluations (Nielsen, 1990).

Experiments showed that individual evaluators performed quite badly. Only between 20% and 51% of the usability problems were found in the interfaces evaluated (Nielsen, 1990). On the other hand, when combining the evaluations together the results become much more accurate (Nielsen, 1990).

Jeffries & Desurvire (1992) have also written a paper regarding usability testing and heuristic evaluation. Their studies showed that heuristic evaluation can work quite well and the method actually found more problems than any other evaluation technique including usability test. Though involved evaluators were trained in usability issues as less knowledgeable evaluators performed poorly. Finally the results of the four expert evaluators were aggregated to achieve good results. In the end the kinds of the problems found by the different usability techniques were quite different (Jeffries & Desurvire, 1992).

Jakob Nielsen’s heuristics describe a specific guidelines that designer or engineer implementing user interface should keep in mind. They are published by Jakob Nielsen in the book *Usability Engineering* (Nielsen, 1993) and on *Nielsen Norman Group* website (Nielsen, 1995) as following:

- **Simple and natural dialogue**
  User should have visibility of system status. The system should always keep users informed of its state. Feedback should happen within reasonable timeline.
- **Speak the user’s language**
  Match between system and the real world. The system should speak the users’ language

with words, phrases and concepts that are familiar to the user. Follow real world conventions, making information appear in the logical order.

- **Minimize user memory load**
  Recognition rather than recall. Minimize the user's memory load by making objects, actions and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be easily retrievable whenever possible.
- **Consistency**
  System should follow standards and consistency. Users should not have to wonder whether different words, situations or actions mean the same thing. Follow platform conventions.
- **Feedback**
  Aesthetic and minimalist design. Dialogues should not contain information which is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.
- **Clearly marked exits**
  User control and freedom. Users often choose system functions by mistake and will need a clearly marked exit to leave the unwanted state without having to go through an extended dialogue. The system should support undo and redo.
- **Shortcuts**
  Flexibility and efficiency of use. Accelerators – unseen by the novice user – may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allows users to tailor frequent actions.
- **Good error messages**
  Help users recognize, diagnose and recover from errors. Error messages should be expressed in plain dialogues (no codes), precisely indicate the problem and constructively suggest a solution.
- **Prevent Errors**
  Even better than good error messages is a careful design which prevents a problem from occurring in the first place. Eliminate error-prone conditions or check for them and present users with a confirmation option before they commit to the action.
- **Help and documentation**
  Even though it is better if the system can be used without documentation, it may be

necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task. List concrete steps to be carried out.

These heuristics are known as "rule of thumb" and are the 10 most general principles for interaction design (Nielsen, 1995).

3.2.2 Usability testing

Usability testing is technique used in user-centered design for evaluating a system or a product by testing it on users. The goal is to uncover the strengths and weaknesses in the usability of a system, product or service. Sometimes usability testing is referred to as *think-aloud testing* because of the importance of communicating participant's' experience and feelings to the usability test organizer (Dumas & Salzman, 2006).

Usability test should focus on the usability of the product and is not useful to answer marketing related questions, like how many products customers will buy. It is advised that the participants are end users or potential end users of the product. This trait makes a difference between usability testing and usability inspection methods, where end users are not included (Dumas & Salzman, 2006). There are multiple ways to perform a usability test. It is possible to ask just opinions of the users and try to see whether they understand the idea of the product or not. In turn the service or the product that you offer to your customers has a specific goal or solution that should solve customer's problem. It is not always easy to find correct demographic group of users to test the product, therefore it is good for the test organizer to specify tasks that will be performed by the end users. The task selection should be done based on the product strengths, weaknesses, or on product traditional or critical use case. When the users are performing their tasks, they should be encouraged to think aloud (Nielsen, 2012). If the participant will be asked to tell his feelings later afterwards, it could be already late and the participant himself will not remember all of the issues that the interface caused. The data must be collected during the test; preferably whole usability test should be recorded on a video. Later it is much easier to review the video to get more precise information. Collected data includes typically qualitative measures, like satisfaction ratings and quantitative measures, like task success and error rates. It is recommended to perform multiple user tests on with same user interface and task cases. The more users will perform the test, the better view you will get on the critical issues of the system or product. Jakob Nielsen points out, that typically five to six users are enough to determine majority of the issues in a system (Nielsen, 2000).

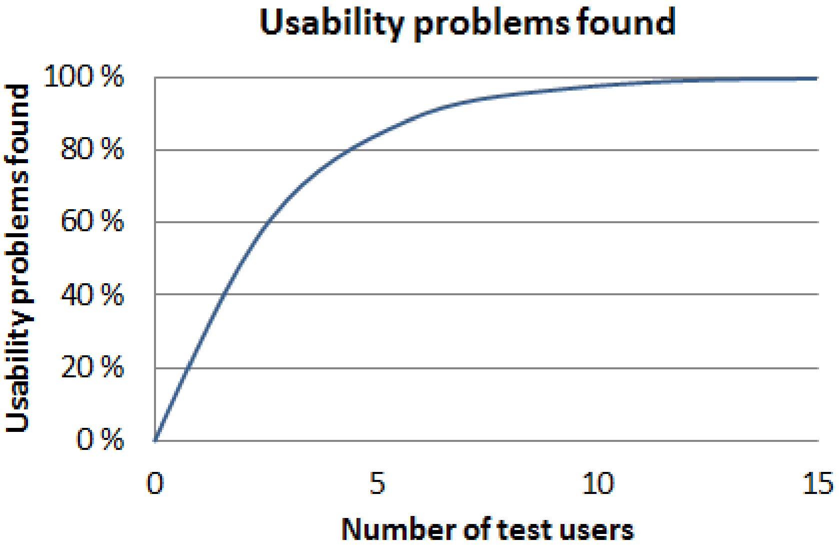


Figure 2: Why you only need to test with 5 users (Nielsen, 2000)

The first five users give invaluable feedback. According to Jakon Nielsen, a single test user will give good insights and will be able to spot almost one of the third issues in your product (Nielsen, 2000). The second user will discover some of the similar issues as the first one, but will be able to provide data that has not yet been found. The second user together with the first one will discover roughly 50% of the usability issues in a product. The third user will find out more issues but not as much in relation to the second and the first user (Nielsen, 2000). As you add more and more users, you will learn less and less due to the fact that after the fifth user, new users will be discovering same issues that the first 5 did (Nielsen, 2000). Figure 2 shows that all issues are only found if tested with 15 users or more. Why Nielsen then does not recommend testing with 15 users? Answer lies in the budget. It is better to spend money and resources on three usability tests rather than spending everything on one usability test. It is good to remember that if your product is intended for different user groups it is highly recommended to have separate tests for each of that group (Nielsen, 2000).

After conducting the usability test, do not forget to communicate the results and recommendations for improvements to appropriate audiences – such as designers, programmers and product owners.

3.2.2.1 Usability testing stages

Early application of usability testing is now considered the ideal (Dumas & Salzman, 2006). In the early days, usability testing was though as a quality assurance testing. When waterfall models were in wide use, the testing itself was performed at the end of the product development cycle. As usability testing was considered to be similar to quality assurance, the User Interface (UI) tests were performed in really late stage. This resulted in poor user interface quality, and because the issues were uncovered too late, there was no time to fix them (Dumas & Salzman, 2006).

Early concept testing tackles early mentioned issue, where the goal is to test alternative user interface design concepts and evaluate them in an early stage. In the very beginning of the product development the concepts are tested with the paper prototypes and easy to use sketching tools (Dumas & Salzman, 2006). The earlier it is possible to include real users in testing the prototypes, the faster development team can react and fix major usability issues. Management rarely understands how costly some of the user interface changes can be, especially if done in late stage.

Testing that is conducted throughout the product's development is called formative (Dumas & Salzman, 2006). The formative test is aimed to guide the design and focuses on usability strengths, weaknesses and improvement. Testing that is executed near the end of product development is called summative (Dumas & Salzman, 2006). The summative test focuses on measuring product's efficiency and its improvement.

The testing process becomes more quantitative and structured as the product design shapes up. In late stages diagnostic testing, and benchmarking and comparison testing play big role. With diagnostic testing the aim is to cover the core functionality and probe problematic areas. Typically tests consist of ten to twenty tasks with the session length of roughly one and a half hours (Dumas & Salzman, 2006). Commonly diagnostic testing includes six to eight users and measures values such as average task completion rate or number of assists from the administrator (Dumas & Salzman, 2006).

Benchmarking and comparison test has a measurement focus for benchmarking product's usability or comparison of different product versions (Dumas & Salzman, 2006). Compared to diagnostic tests, the benchmarking and comparison tests tend to have more of a research experiment approach. These performance-based tests require much larger user sample. When measuring for example time spent on a task, the user sample must be large enough. This is required in order to diminish the skewed results or so called *wildcard effect*[2] (Gray & Salzman, 1998). Individual hotshots that perform some of the tasks much faster than normal person can skew the results in better direction

[2] "An important concern to be aware of in the between-subjects design is the situation in which one of the participants in a group is especially good or bad at performing tasks; Gray and Salzman (1998) called this the wildcard effect." (Dumas & Salzman, 2006)

than in reality it is. The comparison-based test has its own challenges. When comparing two different product designs, it is strongly advised to use two different user groups as there is a risk of contamination from product to product (Dumas & Salzman, 2006). How to find similar or identical user groups? That is really difficult, thus the sample size must be large enough to eliminate these deficiencies.

The most difficult part in conducting benchmarking or comparison tests is the test design itself. The test arrangement should provide measurable and valid comparison between products. Tasks should be as similar as possible and the test administrator's interaction with users should not favor either product. Additionally, it is important to keep in mind how many participants are enough to distinguish a statistically significant difference.

3.3 A/B testing theory

A/B testing also known as split test is a randomized experiment where two variants are used to determine a better option based on the selected metric (Gofman, 2007). Split test is used largely in marketing and web development (Burk, 2006).

A/B testing is often used by marketers to test ideas and drive return on investment (Burk, 2006). Because marketers are always looking for an advantage, they strive to improve market share and win customers. A/B testing method is widely used by marketers to measure success of design, promotion or special offer. The classical example is direct mail where company splits their mailing list and sends out different versions of the newsletter to different recipients (Burk, 2006). In this paper mostly the usability and conversion rate aspect of the A/B test will be covered with focus on web environment.

While usability inspection methods still remain effective practice to evaluate and improve usability of a product, many companies have introduced A/B testing as a support tool for reaching their desired goals. A/B testing has been a popular tool among large business companies but it is starting to get a foothold in small and medium-sized businesses as well (Siroker & Koomen, 2013) (Visual Website Optimizer, n.d.). Even Barack Obama and Mitt Romney had teams of dedicated people working on A/B testing their campaign websites during the presidential race in 2012 (Siroker & Koomen, 2013).

As of writing this paper the A/B testing phenomenon is still quite young and even though some of papers exists it lacks great amount of scientific research which makes it a challenging subject. Search results with keywords like "split test" or "A/B test" will yield non-relative results from

Google Scholar or any other scientific database. It is possible to find a few books from Amazon with search string “split test” or “A/B test”, but the amount of results are still quite low. Additionally most of the written material and discussion found in books and on the web relates to the landing page optimization. In online marketing the term landing page is used to describe a web page where user lands first after navigating to a specific domain (Gofman, 2007). Landing page optimization is a technique where site owner tries to perfect its site through different optimization steps, one particularly popular is A/B testing (Gofman, 2007). Though landing page optimization uses A/B testing practice to improve the general customer experience, bounce rate and conversion from users to customers, it mostly concentrates only on landing page optimization and usage of ready-made tools to achieve the results. There are multiple tools available that will do most hard work for you (Gofman, 2007). Automated A/B tools will make necessary code changes to your site and they will calculate even the significance of the results and provide you with the right suggestion. Only a few clicks are required to make modifications to the site. This is great for regular sites but does not remove the required work for A/B test implementation in greater software projects. In most cases ready-made landing page optimization tools won’t integrate easily when considering something bigger than a standard website.

3.3.1 Implementation practices

In basic A/B testing scenario you introduce two different versions of design and see which performs the best (Burk, 2006). With dynamic websites and functionality it is easy to create an A/B test and show different content or design to different visitors. Each test should be its own independent sample as multiple hypothesis testing can be error prone (Benjamini & Yekutieli, 2001). Thus it is not recommended to run multiple tests at the same time that overlap some aspects of the functionality or design. If multivariate testing is needed, then extra carefulness is required when conducting analysis. Extracting results from complicated test setup can be challenging. Testing only two designs at once helps to analyze and evaluate the results much quicker and easier. Additionally, we can be sure that there is no correlation or dependence on either variable.

Visual Website Optimizer and Optimizely are tools that will make it easy to analyze and calculate the statistical significance of the test (Quora, 2012). For more complex approach there are tools like Mixpanel and KISSmetrics that will work better with Software as a Service (SaaS) systems (Lofgren, 2013). Tools for SaaS systems will require more work and person conducting the test usually needs to plan the test setup and process the results either manually or using automated approach on his own.

Before starting any A/B test, it is essential to understand what needs to be measured. For example, one metric to measure can be conversion rate of registered users. A simple case could be two different websites with different layouts: one which highlights registration form more strongly than other. Measured metric is number of registered users through website A and website B. Keep in mind that poorly chosen metric can usually be the pitfall of the A/B test (Sumner, n.d.). Consequently, it is important to pursue A/B testing alongside other metrics like web analytics, customer satisfaction surveys, average sales etc. (Sumner, n.d.). These metrics provide assurance that optimizations for particular case will not affect negatively other important outcomes.

A challenging question that people ask a lot is: how long should one run the single A/B test (Cohen, 2009)? The rule of thumb is to choose a predefined time range or number of measurements (N) and keep the test running exactly that long (Salmoni & Gupta, 2012). This way we can eliminate the bias in the test. Consider following setup: you start the test and keep calculating significance of the result after every new data point is added. After the result becomes significant you stop the A/B test. This result could be easily biased, if for some reason there were much more entries of version A in the first hours, which made the results significant and thus skewed. Another issue that can arise from the A/B test is too small sample size. When using Pearson's chi square test to analyze the results, minimum sample size is required before we can apply correct test results. Typically authors of textbooks indicate that the satisfactory approximation is achieved when expected sample size exceeds ten or more entries in each cell (Roscoe & Byars, 1971). The statement is only a rule of thumb and should be taken with a grain of salt. Depending on the division algorithm of A/B test it is highly recommended to have a least three-digit number in each cell to avoid margin of error.

The biased approach of calculating significance of the result is not the only incident that can skew the outcome of A/B test (Salmoni & Gupta, 2012). There are actually many variables that can cause incorrect results, such as holiday season, time of year or a big press release. The A/B test is not unambiguous and should be analyzed carefully. In order to confirm the results one can arrange two same A/B tests during different time period. If the results will yield same conclusion it can be assumed more safely that there were no external factors skewing the outcome.

After metric and time range for the A/B test has been selected the next step is to decide how and where to save all the data that will be gathered. We highly discourage implementing your own system and tools for collecting data points. Firstly it is much more time consuming to create really good and scalable system that can handle high traffic spikes than you might first think. Secondly the analysis tools and data point query language should be easy, so also non-programmers could extract relevant information with ease. Company should never do things that are not company's main competence (Foss, 1997, pp. 235-237). There are a lot of ready-made tools on the market already with a reasonable pricing. Especially if the goal is to optimize your company website, for simple website A/B testing, there are tools like Optimizely (http://optimizely.com), Visual Website Optimizer (http://visualwebsiteoptimizer.com) and Unbounce (http://unbounce.com). With these tools it is possible to optimize your website without the coding knowledge. Using *What You See Is What You Get* (WYSIWYG) editors it is possible to change your site layout completely with just a few clicks. After two versions of the site are created the next step is to copy paste JavaScript code snippet into your website code. The code snippet will do all the work, like splitting the traffic into two samples (version A visitors and version B visitors), calculating the significance of the results and showing real time statistics. If you plan to create your own A/B test logic and plan to test the website or a Software as a Service (SaaS) that can be accessed by search engines, it is recommend to use caution as A/B test can disorder Search Engine Optimizations (SEO) (Nassimian, 2011).

3.3.2 Amazon case description

It is considered that it was Amazon who pioneered the A/B testing in e-commerce (Eisenberg, 2008). Over 76 million people have purchased from Amazon.com and still counting (Steiner, 2008). With these kind of numbers even a small increase in conversion rate, like 0.1% can mean huge difference in the business. Amazon knows this and that is why it has been fine tuning its shopping cart for many years.

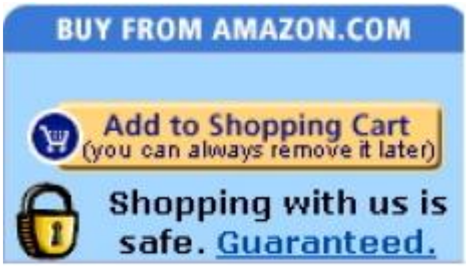
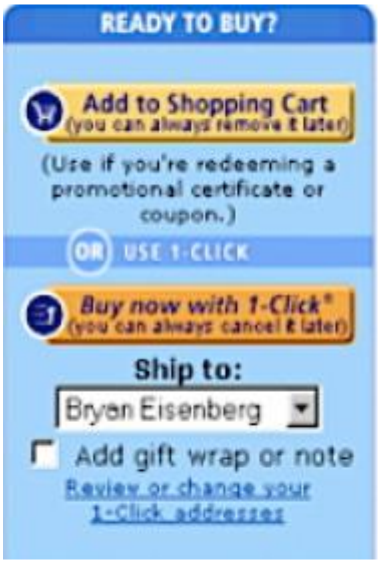

BUY FROM AMAZON.COM
Add to Shopping Cart
(you can always remove it later)
Shopping with us is safe. Guaranteed.

Figure 3: Amazon's one of the first add to cart buttons (Eisenberg, 2008)

As seen from Figure 3, it is from the yearly days of internet and e-commerce when people were still cautious of online shopping. With the text indicating that "you can always remove it later" and "Shopping with us is safe. Guaranteed" Amazon assured users that the shopping is indeed safe and it is not the end of the world if you click wrong button.



(Use if you're redeeming a promotional certificate or coupon.)

Figure 4: Amazon cart “buy now with 1-click” (Eisenberg, 2008)

In Figure 4 “Buy from Amazon” now changed to more inviting “Ready to buy?”. The original buy button stayed the same, but Amazon added their new feature “1-click buy”. They even used different color to bring out and distinguish the “1-click buy” action to increase impulse shopping.

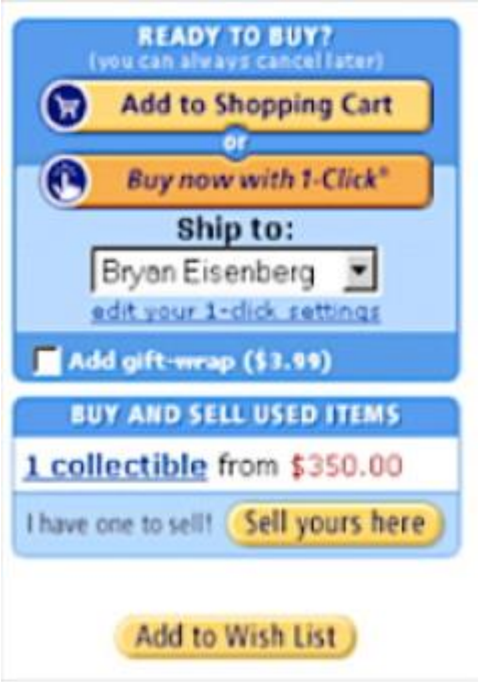


Figure 5: Amazon new add to cart layout (Eisenberg, 2008)

Amazon did some facelift and font changes to their shopping cart as seen in Figure 5: removing large amount of text, shrinking it and placing under “Ready to buy” title. Also bringing two buy buttons closers to each other to create more condensed look and point out that the two buttons have same final goal.

With these changes, there was also interesting story that Bryan Eisenberg (2008) described in his article:

> *"The funny thing that happened when Amazon made these changes was that many of our clients at the time decided they should also remove point-of-action assurance from their Add-to-Cart buttons. We told them it would hurt their conversion if they changed it — and, sure enough, against our advice, the clients changed it and conversion dropped. Yet Amazon kept the new buttons. So the question remains…*
>
> *Why would they switch to buttons that don't convert as well?*
>
> *Because conversion isn't the only metric that matters. If you look closely, you'll notice they made the "Ready to Buy" area take up about half the space of the previous version. Why? Because they quietly launched a marketplace to resell used goods, deciding it would boost profits if they didn't have to stock and ship everything themselves — a fundamental shift in their business model."*

The moral of the story is: do not copy other people or company ideas if you are not fully aware of the business issues and strategy involved. Test and experiment with your own goals and business issues.

3.3.3 Analysis method

A/B testing analysis can be done with two statistical tests: the Student t-test and Person's chi-squared test. Student t-test determines if two sets of data are significantly different from each other and usually used only when the sample distributions are normally distributed. Person's chi-squared test, which is one of the variations of chi-squared tests, is used to assess two types of comparison: test of goodness of fit and test of independence (Plackett, 1983).

In our case the sample is not normally distributed (Cohen, 2009). If our measurement metric is registered users through a website, which means that possible results are discrete 1 or 0, depending whether user registered or not. This indicates that this is a two-valued discrete distribution. Student-t test can work also when the samples are not normally distributed on condition that sample size (N) is large enough.

The Person's chi squared test null-hypothesis in our case is that the A/B test results are due to chance alone. Our goal is to reject that hypothesis and prove that user actions are meaningful.

The formula for calculating the test statistic is following[3] (Plackett, 1983) (Cohen, 2009):

$$X^2 = \sum_{n=1}^{m} \frac{(O_i - E_i)^2}{E_i}$$

Where:

X^2 = Pearson's cumulative test statistic which follows X^2 distribution

O_i = observed measurement

E_i = expected measurement

m = the number of samples

In the simple A/B test case, where we have only two variants: $m = 2$. Our hypothesis is that results are due to chance alone. Both variants have 1/2 chance, thus expected values are $E_i = n/2$, where $n = O_1 + O_2$.

Changing $O_1 = A$ and $O_2 = B$, where A is to be larger number of the observed values, we get following formula:

$$X^2 = \frac{(A - \frac{n}{2})^2}{\frac{n}{2}} + \frac{(B - \frac{n}{2})^2}{\frac{n}{2}}$$

$A + B = n$, thus the square difference of $\left(A - \frac{n}{2}\right)^2$ is the same as square difference of $\left(B - \frac{n}{2}\right)^2$

Prove and calculations that both square differences are same, when $n = A + B$

$$\left(A - \frac{n}{2}\right)^2 = \left(A - \frac{A - B}{2}\right)^2 = \left(A - \frac{1}{2A} - \frac{1}{2B}\right)^2 = \left(\frac{1}{2A} - \frac{1}{2B}\right)^2 = \frac{1}{4A^2} - \frac{1}{2AB} + \frac{1}{4B^2}$$

$$(B - \frac{n}{2})^2 = (B - \frac{A - B}{2})^2 = (B - \frac{1}{2A} - \frac{1}{2B})^2 = (\frac{1}{2B} - \frac{1}{2A})^2 = \frac{1}{4B^2} - \frac{1}{2AB} + \frac{1}{4A^2}$$

thus the formula can be reduced to:

[3] Variable names have been changed to be more figurative and descriptive.

$$X^2 = \frac{4D^2}{n}$$

Where: $D = A - n/2$

The use of a chi-squared distribution requires the use of degree of freedom. In an identical manner as with the Student's t-distribution the sample size determines which distribution to use. If the sample size is m, then there are n – 1 degrees of freedom (Lane, 2013). Because the sample size in our case is two ($m = 2$) and there is only B which depends on the variable A (or vice versa), there is only one degree of freedom. Thus looking at chi-square distribution it is required to exceed[4] critical value of 3.841 to achieve confidence level of 95% (NIST/SEMATECH, 2012). For confidence level 99% it is required to exceed critical value of 6.635.

For the simplicity of the formula, we select $X^2 = 4$ as a threshold for statistical significance. This means that if we exceed value of 4, we are well beyond the 95% confidence level. Solving D^2 the formula becomes much easier to use:

$$4 < \frac{4D^2}{n}$$

$$D^2 > n$$

Where $D^2 = \left(A - \frac{n}{2}\right)^2 = (A - \left(\frac{A+B}{2}\right))^2 = \left(A - \frac{A}{2} - \frac{B}{2}\right)^2 = (\frac{A}{2} - \frac{B}{2})^2 = (\frac{A-B}{2})^2$

$$(\frac{A-B}{2})^2 > n$$

To achieve over 95% statistical confidence level in simple A/B test, one is required to have greater square difference of results divided by two than the number of trials.

It is essential to remember that the chi-squared test must be conducted with a sufficiently large sample size. Commonly it is considered that sample size should exceed ten or more frequency points in each cell (Roscoe & Byars, 1971). This means that in both A version and B version you should have ten or more measurements or action points. If the chi squared test is conducted with a smaller sample size, the test will yield an inaccurate conclusion. The described behavior is called a

[4] Because we are trying to reject the null hypothesis, which states that the results are due to chance alone, we need to exceed the critical value (NIST/SEMATECH, 2012).

Type II error (Shermer, 2002). In type II error analyst accepts the null hypothesis when he should be rejecting it or vice versa.

3.3.4 Conversion rate and optimization

Conversion rate is a probability that user's initial action will ultimately yield a desired goal – a sale for instance (Kanich, et al., 2008). A typical example is a website where user who ends up on the site will actually register for a service and becomes a member.

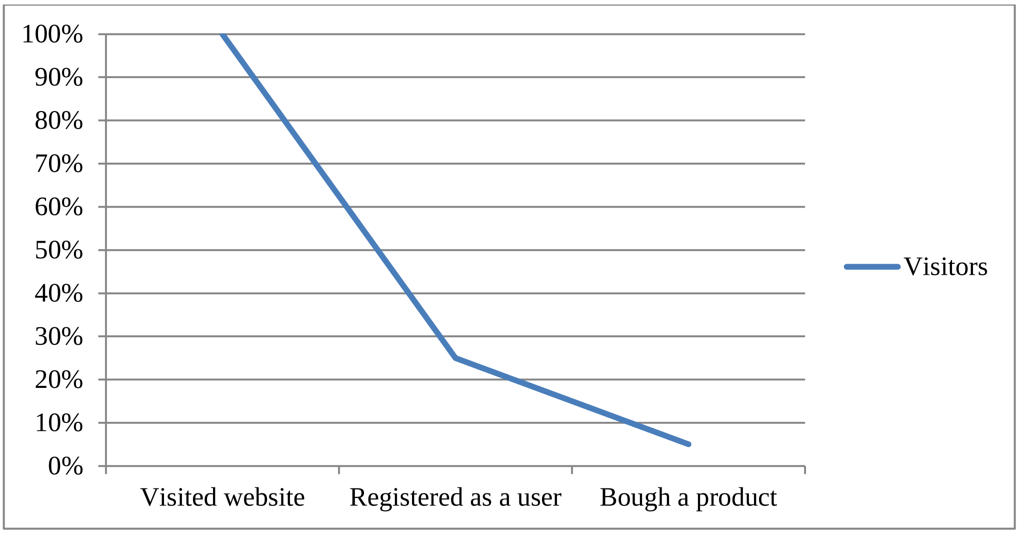


Figure 6: A basic conversion rate graph

From Figure 6 can be seen a basic example of conversion rate graph which illustrates conversion rate from website visitors to registered users and customers. The starting point and first step is a visit to the site, which is 100% of the visited users. Then the next step is a registration, where only 25% of visited users convert to and lastly a bough product which only 5% of the visited users convert to.

Mostly A/B testing is done to increase conversion rates. Many companies practice it and some of them do it even for a living: conversion-rate-experts.com. To properly increase conversion rate of your website or service it must be optimized throughout all steps. For example improving your marketing in Facebook and Twitter does not necessarily increase your conversion rate if your links from Facebook and Twitter point to non-optimized landing pages where customer gets lost.

Conversion rate and conversion funnel are really closely related terms and sometimes they are used interchangeably. The term funnel is used to describe the decrease in numbers that occur after each action or step of the process. The metaphor of a funnel is used because the process is frequently delineated as a pipe with a wide mouth or inverted pyramid (Kanich, et al., 2008).

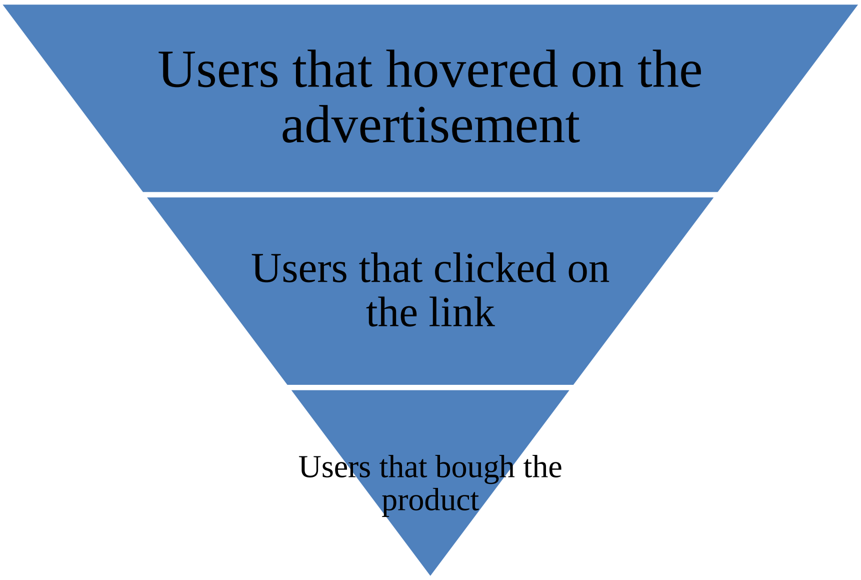

Users that hovered on the advertisement
Users that clicked on the link
Users that bough the product

Figure 7: A general presentation of the conversion funnel

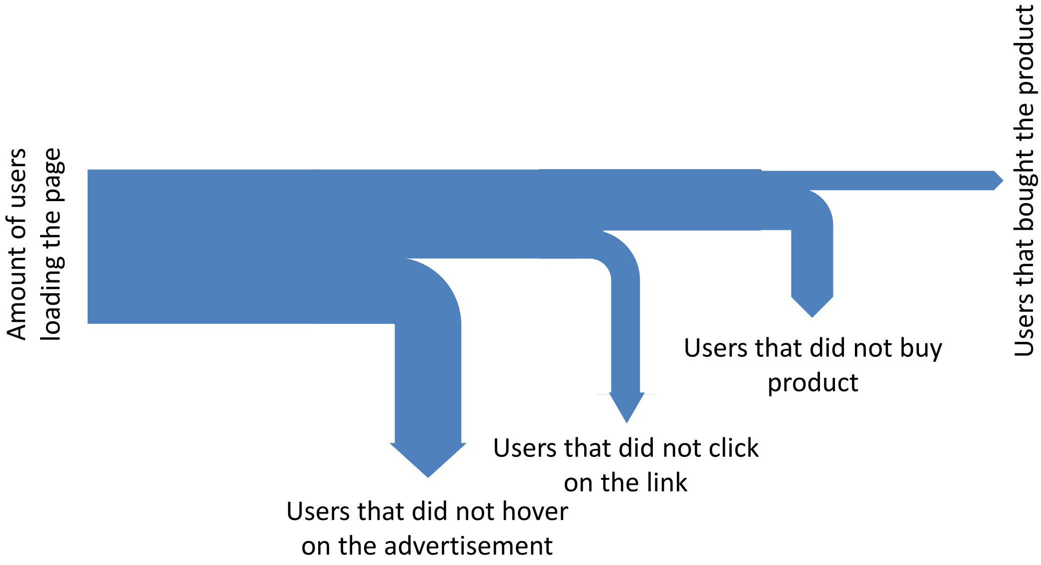

Amount of users loading the page
Users that bought the product
Users that did not buy product
Users that did not click on the link
Users that did not hover on the advertisement

Figure 8: Another representation of the conversion funnel

The funnel presentation in Figure 7 is a general overview of the process, while in Figure 8 it can be seen from the thickness of the bars more precisely the amount of users dropping out in different steps. These are just two of the most common approaches to represent the conversion funnel as when system and analytics get more complex the funnel presentation can have even multiple funnels connecting to each other (Google, 2013).

Challenging aspect of A/B testing is that an organizer needs to know which metrics are important in that particular situation. Furthermore the organizer must have a deep insight of the system to understand which element should be tested to enhance that metric or key value. To achieve better conversion rate in one funnel, tradeoffs in another funnel need to be made. Consider following example:

A webshop owner plans to A/B test his `add to cart` button with the original `add to card` button and a new `want` button. Version A is single `add to cart` button while version B is an `add to cart` button with `want` button next to it. After a few weeks of testing he sees that `add to card` button receives 5% less clicks if it is paired with `want` button. The `want` button receives double the amount of clicks than `add to cart` button. A webshop owner makes a hasty decision and drops the `want` button to keep his sales up. Was the decision correct?

The answer depends completely on his business model and company goals. Although making that kind of decision could prove wrong. If the goal was to increase the sales in the long run, the want button could provide him with marketing. Assume that wanting a product on that website would share it in social media. People clicking want would be advertising the goods sold on that website across the globe for free. In the long run the marketing trick could pay off and increase the amount of regular clients and correspondingly profit.

3.4 Comparison between A/B test and usability assessment methods

A/B test and usability assessment methods are completely different methods for different purpose but they both still have similarities and common overlays. Usability is rarely tied to revenue and corporate wellbeing, but rather concentrates on the users and user experience. A/B test is marketing oriented and drives test with revenue based aspect in mind. Even though the goal of usability assessment methods is to improve usability of a product and user experience of the user, the long-term business goal is usually improvement in sales and getting an edge over the competitors. The long-term goal with the A/B testing is increasing sales as well and converting potential customers to buying customers. The small difference in two methods is that generally with A/B testing company aims to improve conversion to action with existing customers, while usability assessment methods emphasizes on early testing of the system (Bock, et al., 2005). Usability assessment methods focus on improving existing usability of the product as well, but try to facilitate early focus and integrate the process into development, especially before the implementation of functionality and user interface has begun (Holzinger, 2005). These roles may change and are changing even now, only time will tell will there be some kind of better combination of usability tests and A/B testing to increase the revenue – or even some kind completely new approach.

As pointed out by Jeffries and Desurvire, usability assessment methods require that the evaluator is experienced in his field as inexperienced evaluators provide usually poor or incorrect results (Jeffries & Desurvire, 1992). A/B testing is far more forgiving in that matter and only requires implementation of two versions that are going to be tested. Although knowledge of the assessment methods may be helpful when deciding which part of the system to put into A/B test.

3.4.1 Strengths and weaknesses of A/B testing

A/B testing is an effective method that can provide good insight if done correctly. Taking into account all the pitfalls, like minimum number of sample size, time period for measurement and correct tool to analyze the results will yield accurate results. It is good to remember that the correct results will not always provide definite answer; usually A/B testing requires a lot of trials and errors (Farakh, 2013).

With A/B testing approach the organizer is responsible of creating the test and choosing the elements that he/she wants to test. It seems that many people and sites perceive that A/B testing is just a magic tool that will increase conversion with just very little effort. The internet is full with the stories about successful A/B tests, where changing just a small portion of the website resulted in huge increase in conversions (Visual Website Optimizer, n.d.) (Gofman, 2007). Generally many trials are required to get meaningful results and successful A/B tests are a rare thing (Farakh, 2013). It can really take time before you hit some meaningful result and when you do – the increase in conversion most likely won’t be 200% but rather around 5% or 10% (Conversion Rate Experts, ei pvm). Conversion rate can increase 200% or more, but that just means that there was something really wrong with a previous design. If the company has some kind of usability expert or design team working on the site or the product, you will probably never see more than 25% increase in conversion rate in a single test (Conversion Rate Experts, n.d.). Sometimes A/B testing can feel more like shooting a gun with a blindfold rather than success after success. The reason for that lies in nature of the A/B test which will not give you any proper user feedback about your product or website. Looking at logs and statistics can give insights, such as: which stage of the funnel users drop out or that version A of the page is performing a bit better than version B. The weakness of the A/B testing is that you may never know for sure what causes that kind of behavior. Only testing small part of the product or a website one by one can be more precise and descriptive. For example changing only a title of a heading can tell for sure that one title was better than another and the increase was due to that title name change. This implies that conducting a large umbrella tests that cover large part of the system will not yield a definite answer to the cause.

A/B testing requires more data than any usability assessment method and may not be appropriate for low traffic sites (Gofman, 2007). This can be an issue for companies that do not have enough traffic to generate meaningful results. This scenario applies especially for startup and small companies which haven’t yet established strong foothold in their field. Additionally if company’s customer base is quite small, the conversion increase of 5% will not seem to matter that drastically. Compare for example a small company with 100 clients versus enterprise company with 1 000 000 clients. Even that 1% sale increase can pay multiple employee salaries in large companies.

What makes A/B testing so powerful and useful tool? Unlike usability assessment methods A/B testing is done in real environment with actual users. It is difficult to get any closer to real numbers and results than A/B test. Usability tests can be done remotely and online as well, but user is still aware of the situation unlike in A/B test case. Dissimilar to usability test, the A/B test organizer does not need to search for suitable candidates for the experiment but can concentrate on the test itself.

3.4.2 Strengths and weaknesses of usability assessment methods

Each usability assessment method has its own strength and weakness but all of them are struggling with finding suitable candidates for the test as users or in heuristic evaluation finding usability experts (Nielsen, 1994). Usually in heuristic evaluation it is recommended to have at least a few usability experts evaluating your product – the good ones are generally not easy to find and definitely not cheap.

Unlike in A/B testing with usability assessment methods it is much easier to find starting point of the problem and to get a real user or expert feedback on the issue. Expert analysis can give results without your own input or work as it can be done by just sending the product to the evaluator. With usability tests and cognitive walkthrough it is contrariwise. The organizer needs to engage with users and take his role with required expertise.

Additionally in usability test participant is monitored by evaluator. Evaluator can make participant either comfortable or uncomfortable which in turn can increase the stress of the user thus skewing the experiment flow and results (Rubin & Chisnell, 2008, pp. 211-213). Furthermore the test and results largely depend on the evaluator as well (Hertzum & Jacobsen, 2001). Inexperienced organizer can unknowingly distort the results by providing uneven assistance for the test participants. Also organizer's body language and the atmosphere of the laboratory can affect the performance of the users. If the organizer himself is stressed about the test, the participating user will most likely be stressed too. The main task of the evaluator is not to just gather the data and evaluate it, but try not to effect participants' opinions and still encourage the users to think aloud and comment on the features of the product. This is what makes a good evaluator stand out of the rest. Studies have shown that different evaluators evaluating the same system with the same usability assessment methods detect substantially different types of usability problems in the system (Hertzum & Jacobsen, 2001).

Studies have also showed that using only one usability assessment method does not find all the issues in a system (Nielsen, 1994). Certain problems are overlooked by some usability inspection methods (Nielsen, 1994).

What makes usability assessment methods stand out is that they can be practiced in any situation. It does not need to be only a website or software. The evaluated system can be an actual physical product. Finally with usability assessment methods only handful of candidates is needed to assess a product.

3.4.3 Comparison overview

Both A/B testing and usability assessment methods are powerful tools and they are suitable for certain situations better than the other. Even sub-methods of usability assessment methods have their own strengths and weaknesses. Especially usability testing is widely known and used practice that is often compared with A/B testing.

Usability testing and usability inspection methods are powerful tools but require generally more resources. Hollingsed and Novick (2007) indicate that evaluation or usability testing can prove to be expensive and time consuming. A/B test in turn can be costly when implementing a new variation design from scratch. With usability assessment methods evaluators can evaluate wider area of the product – while A/B testing results can only relate to the small specific part of the product that is being A/B tested. A/B test may not provide a clear reason for user behavior – as Gerken, et al. (2008) points out logging analysis do not answer all of our usability questions and in some cases even raises new ones. On the other hand A/B testing is done in real environment and users do not even know about the test itself and act as they would in a normal use case. Usability assessment methods rely on a lab or on-site experiment in a controlled environment.

It is not always easy to decide what should be compared or tested with A/B test (Sumner, n.d.). Heuristic evaluation or usability testing provides concrete feedback regarding the issues within a product. Evaluator only needs to know the use cases for the product to evaluate it using usability assessment methods.

Usability assessment methods are more suitable for situations where there is not enough traffic or customers to be practicing A/B testing. Usability assessment methods require only a few candidates to reveal most of the issues in the product while A/B test rely on quantitative results.

	A/B testing	Usability assessment methods
1)	Live testing	Immediate results
2)	Quantitative metrics	Qualitative as well as quantitative metrics
3)	Can measure small performance difference	Versatile use
4)	Low maintenance	Behavioral insight

Table 1: Benefits of usability testing and A/B testing (Mesibov, 2011) (Nielsen, 2005)

Compared to other methods, A/B testing has four advantages as can be seen from Table 1. It measures actual behavior of customers under real-world conditions (Nielsen, 2005). It can measure very small performance difference with high statistical significance and can resolve trade-offs between conflicting guidelines (Nielsen, 2005).

Usability assessment methods can generate results with as little as partially-complete design, wireframe or prototype (Mesibov, 2011). With usability testing metrics such as task duration and completion rate can be calculated, but also can answer why users take certain actions. Usability assessment methods can be versatile and do not need to be limited to one measurable metric. For example A/B testing tends to be limited to measurable actions such as: sales for e-commerce site, users subscribing to a newsletter, users downloading a white paper. Usability assessment methods and particularly usability testing can be set up as a process to generate constantly updated user feedback and behavioral insight.

4 Research methods

In this empirical research two methods were used to gather data. One was done with quantitative approach and another with qualitative approach. The data of the quantitative research was based on the A/B test and was gathered and analyzed using methods in chapter 3.3.3 Analysis method. The qualitative research comprised of the usability test. The latter test focused on mobile platform while A/B test was conducted based on the findings of the usability test.

To understand the topic of the A/B test better, scientific articles and books related to A/B testing (also known as split testing) were explored. Following keywords were used to find research papers: “A/B test”, “AB test”, “AB testing”, “split test”, “split testing”, “conversion funnel”, “marketing funnel”, “conversion rate optimization”, “e-commerce conversion rate”, etc. Many searches ended with results related to e-commerce and how to convert store visits to purchases using historical data or some other approaches. A/B test related information based on scientific research was difficult to find and did not yield good search results. Mostly A/B related material was found from the web, blogs, corporate sites and other forums.

4.1 Case description

The case study was conducted in the software company named Kiosked where I am currently working. Kiosked is a leading platform in enabling Smart Content. With Smart Content Kiosked turns any online content like images, videos and applications into interactive and viral storefronts (Kiosked, 2013). Smart Content enables relevant and targeted content in media, for example user who is watching a performance of an artist in a YouTube video can see the information about the clothes the artist is wearing and with a few clicks the user can make a purchase of those particular items.

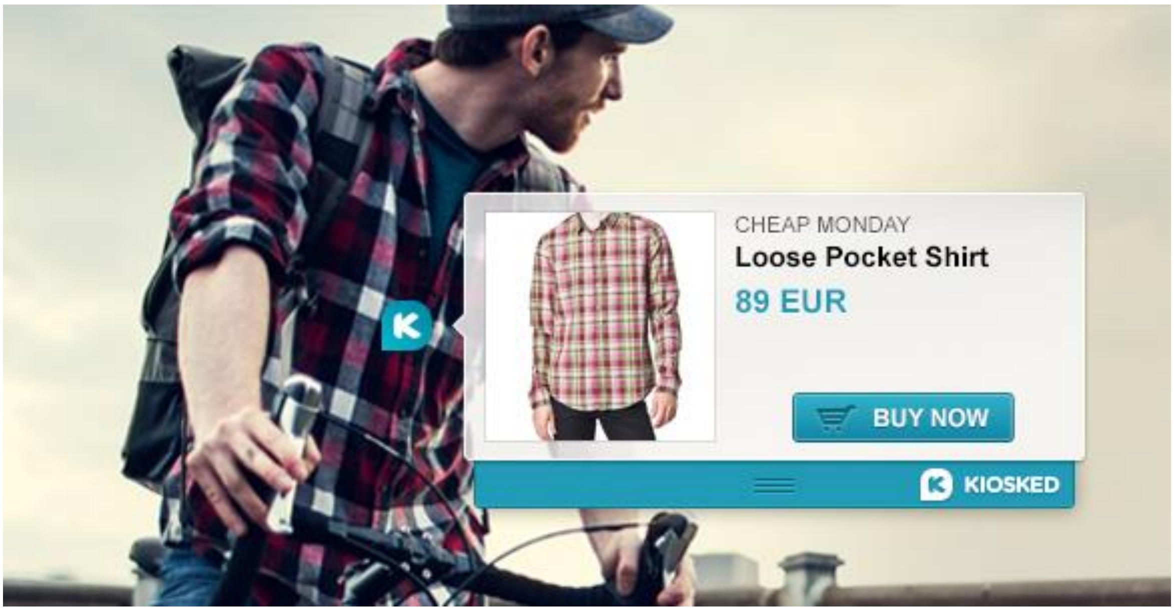


Figure 9: Illustrating the concept of Kiosked. Relevant content placed inside image.

The content can be added inside the media using two different approaches: user tagging[5] the media by himself or leaving that task for the Kiosked system to handle. The system will automatically tag relevant content into media based on the media name and keywords as well as product name, keywords, categorization, site content and user based preferences. When media is tagged with the content we call that action kiosking. Also the media that has been tagged becomes a kiosk, like a kiosk in physical world from where you can get information and buy different products.

By default images and videos are static and non-interactive elements (except for video control elements, like play and pause buttons). To enable kiosks and user interaction in media like images and videos, Kiosked is using JavaScript code. JavaScript is a programming language that was designed to work in browsers so that engineers of web sites or web based services could easily create dynamic and comprehensive user experience. When user is adding a kiosked video or image to their website, he is actually adding element that links to appropriate JavaScript file. The JavaScript file then generates all the necessary elements and events in order for the user interaction to be possible. For example to include a kiosked video on your own site, one needs to copy paste appropriate script tag (JavaScript file) to his own website which looks like following:

```
<script type="text/javascript" 
src="http://widgets.kiosked.com/widget/kvideo/co/794/id/2425620.js"></script>
```

[5] In this case the meaning of tagging is: dragging and dropping the product on top of the media

When customer visits website with the video script tag, the browser of the customer will make request to the Kiosked web servers and will request the JavaScript file which is located at http://widgets.kiosked.com/widget/kvideo/co/794/id/2425620.js. The script tag then will load the video that is assigned to the id 2425620 and after than appends all the necessary kiosks that are linked to that video.

Kiosked supports automatic kiosking of the images as well, where user only adds one script tag to the whole site and all the images on that specific website will get automatically kiosked with appropriate products. Using Intelligent Matching algorithm Kiosked is able to match relevant products with sites content to produce good experience for the customer.

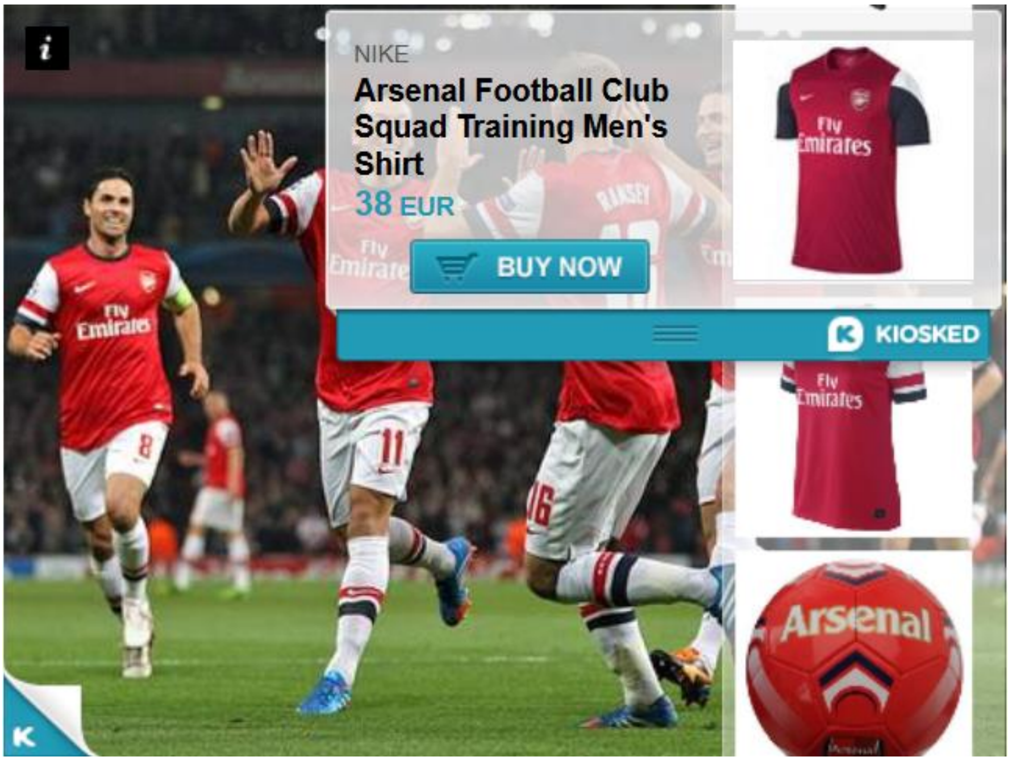


Figure 10: Image on a customer site which has been kiosked with Intelligent Matching.

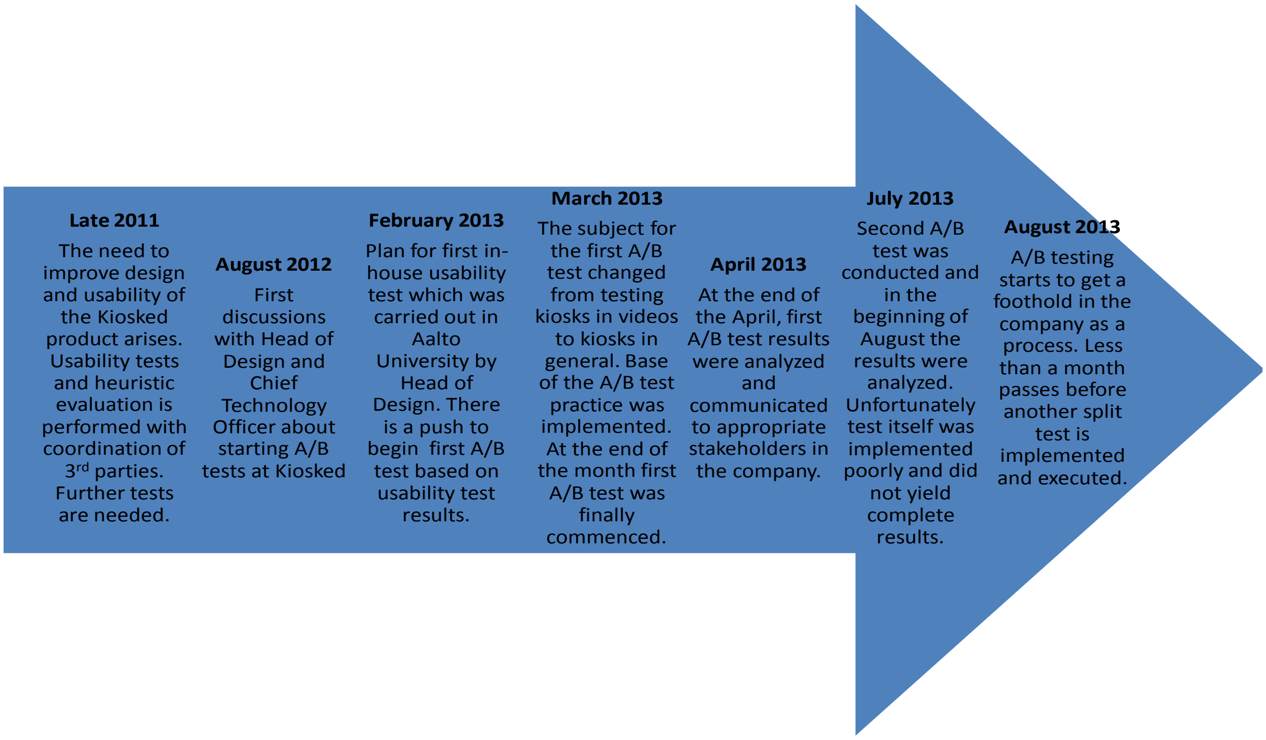
4.1.1 Company background

The company currently employs around 50 employees and its main office is located in Helsinki, Finland but has also offices in London and Los Angeles. Structure of Kiosked is consisting of three major departments: sales, development and management. All of the three departments roughly employ same amount of people. I started in Kiosked in June 2011 and have been working in the company almost three years now. Despite my job description as software developer I have been doing different tasks from network administration to software development as well as executing and maintaining partly of the analytics implementation.

Currently the company is working with Rovio and many other big names to expand and grow its business globally with the focus on United Kingdom and United States.

4.1.2 Conducted research and tests

The first steps of A/B test planning began in late third quarter of 2012 but the design and implementation began in early 2013. The objective of the A/B test was to gather broader business intelligence and get insight of the user behavior in Kiosked Enterprise System. Charlotta Turku who was our head of the design had already performed the qualitative research regarding the user behavior in February 2013. The usability test was based on the mobile version of kiosk and the task was to find out how easily users understood the Kiosked concept and could they actually buy products through the interface. Additionally heuristic evaluation and usability testing by 3rd party was performed on user interface of our main dashboard. The unfortunate part is that both results could not be used in this research as a reference or comparison. The reason for this was that user interface of the dashboard was rebuilt from scratch to correspond updated user requirements thus making the heuristic evaluation results outdated.



Late 2011
The need to improve design and usability of the Kiosked product arises. Usability tests and heuristic evaluation is performed with coordination of 3rd parties. Further tests are needed.

August 2012
First discussions with Head of Design and Chief Technology Officer about starting A/B tests at Kiosked

February 2013
Plan for first in-house usability test which was carried out in Aalto University by Head of Design. There is a push to begin first A/B test based on usability test results.

March 2013
The subject for the first A/B test changed from testing kiosks in videos to kiosks in general. Base of the A/B test practice was implemented. At the end of the month first A/B test was finally commenced.

April 2013
At the end of the April, first A/B test results were analyzed and communicated to appropriate stakeholders in the company.

July 2013
Second A/B test was conducted and in the beginning of August the results were analyzed. Unfortunately test itself was implemented poorly and did not yield complete results.

August 2013
A/B testing starts to get a foothold in the company as a process. Less than a month passes before another split test is implemented and executed.

Figure 11: Process of conducted tests at the company

The first A/B test was performed in March 2013, the second one in July 2013 and the third in end of August 2013. The cause behind so few tests was structural inertia of the company. The A/B tests had to be approved by the product council before conducting any test. This led to really clumsy and slow system which does not work in A/B testing environment. A/B test requires rapid prototyping and testing of the different implementations. Whole idea of the A/B testing is to test all the assumptions that people have so that the decisions are not left for the company's management or one person to decide.

4.1.3 A/B test cases

The A/B tests followed after the usability test. Based on some of the findings from usability test we wanted to know if the opinions of certain users were actually correct. Charlotta found out that users were a bit skeptical about the "buy now" button that you can see in the Figure 10. Some of the users were afraid to click it because they were unsure what would happen. With the current setup clicking the buy button actually takes user only to the website or web shop from where user can order the product through additional steps. The users naturally did not know this and some of them were afraid of clicking it. As pointed out in chapter 3.1 and in Amazon case description it is quite common for people to avoid things that they are not familiar with. In this case users were not familiar with Kiosked service and were unsure of the "buy now" button behavior.

The purpose of the first A/B test was to test mentioned user theory. We came up with alternative button and named it "go to store". We also kept the old "buy now" button and made it version A when "go to store" button was a version B. Buttons were in every way similar to each other expect the text that was written on the button. To make text fit into the button, we had to make "go to store" button a bit wider than "buy now" button.

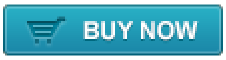


Figure 12: Buy button used in split test (version A)

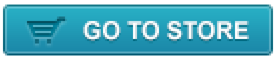


Figure 13: Go to store button used in split test (version B)

The objective of the second A/B test was to see how the kiosk views would convert to purchases if we keep the old "buy now" button but enable the link in the title, also seen in the Figure 10. In kiosk original design the title was always a plain text and not a clickable link. We created two different kiosks where one had a link in the title and another one was a kiosk with the default text in the title. Otherwise kiosks were identical and the link in the title took the user to the same page as the "buy now" button.

The third A/B test was conducted to validate the requirement that came from the product council. The requirement stated that kiosk should show more product information when opened. In the current design user could always access more detailed product information but it required a separate user action. The product council wanted that detailed information to be visible right away when the kiosk is opened without any extra user actions. Our concern was that the flood of extra data would eat up the conversion rate of the buy button. When user is presented with a lot of information at once it could scare him or her from interacting with the kiosk further thus lowering the conversion rate.

4.2 Data collection and analysis

Data was obtained from the different blogs, websites and software that were already using Kiosked service. The data gathered from the visiting users was stored using third party software Mixpanel[6]. The tool provided us with the necessary framework to gather and store the data in easy manner as well as to extract the data in human-readable-form. Mixpanel provides an Application Programming Interface (API) and a library for their clients so they can send HTTP requests to the Mixpanel servers that store the information. Later on client can access web based user interface of Mixpanel dashboard to make queries to obtain the saved information or data points as Mixpanel calls them. To analyze data client can use automatically generated graphs, funnels and retentions to get an insight of the user behavior.

Obtaining the data from the Mixpanel did not yield final results or value for us. We had to calculate significance of the data separately to see if A/B test result was meaningful or not. In calculations we used the same formula as described in chapter 3.3.3 Analysis method:

$$(\frac{A-B}{2})^2 > n$$

Where A stands for results for ‘winner’, B for ‘loser’ and n is total number of results. For example if we would have tested two different versions of the website and the A version would receive 1500 signups and B version 900 signups. The total amount of results n would be 2400. So placing values in the equation we derive following result: $90000 > 2400$. The formula is true, thus the example A/B test result is meaningful.

[6] Mixpanel (https://mixpanel.com/) is an platform for collecting and analyzing data for mobile and web based systems.

The A/B testing was conducted with the help of Charlotta Turku who assisted with the planning. The usability test was performed by Charlotta Turku alone and it was based on the mobile version of the Kiosked software. Twelve users participated in the test and were interviewed. The goals of the test were to find out whether or not users would understand concept of Kiosked and would they buy products through kiosked images.

5 Empirical research

Empirical part consists of one usability test and three A/B tests. The whole empirical part was done in one year. The first A/B test was planned and conducted based on the results that Charlotta discovered prior in the usability test.

5.1 A/B-testing preparations

The A/B test planning started in summer 2012 and at first the intention was to cover only the video aspect of the Kiosked. After months later the plan for split testing video features was discontinued and split testing for kiosks in general was proposed. The first step of empirical research was finally decided in early 2013. Despite the long process the real challenge of the A/B test in Kiosked was finding suitable tools.

5.1.1 Tools used

Most of the popular tools[7] and the sites we found including Optimizely, Visual Website Optimizer and Google Analytics concentrated on the optimization of the standard website, not a web based platform as it was the case with Kiosked. Even though Kiosked platform is a web based system it drastically differs from the standard website and could not be adapted to the tools that were found.

Later on it was discovered that Google Analytics could be actually the solution for our A/B test requirements. Google already included the funnels and the goals in their analytics, so it was relatively easy to create custom goals and monitor the visitors and actions. There was only one small issue – adding the Google Analytics script into our own JavaScript file would mean that any customer that would add kiosked content, like video or image to their site would be adding our Google Analytics script on their site as well. This would give us the entire user statistics from the customer site too. As there was no option to disable this behavior we decided to avert that approach – imaginably this kind of feature would not make any customer happy.

After the disappointment with Google Analytics there was devotion and eagerness to implement custom made solution for A/B test. Logging of the actions and events is actually simple to implement. The challenge lies in making the system scalable. One server can handle the load when talking about a few requests per second. The situation changes when enterprise level system receives hundreds or even thousands of requests per second. The A/B logging system should be able

[7] All the tools that were considered are: Optimizely, Visual Website Optimizer, Google Analytics, Flurry, Kontagent, Mixpanel. Only Mixpanel was chosen as other relied either on standard website or mobile based analytics

to handle the load generated from the users. Furthermore only logging user actions into a system does not solve nor does it generate meaningful analysis of the data. It is obvious that it is a job of an analyst to process the data and draw a conclusion from it, although making the data more accessible and processable will help the analyst in a long run. Analyst should have an easy access to whole data and it should be simple to query. When dealing with big data we are speaking of thousands actions per day. That data must be filtered and processed preferably with automated tools before analyzing it further. After realizing amount of work it requires implementing a robust system for only A/B testing features we gave up the idea as it is not Kiosked main competence to create logging analysis software.

In the end we found a suitable tool for conducting our A/B tests: Mixpanel. Mixpanel enables customer to create custom events, funnels and send in one event arbitrary amount of data. Additionally it is possible to send only the data that is really needed. This was a game breaker for us as we did not want to collect our customers' website statistics. In the end we managed to include the Mixpanel library in our script in a way that did not interfere with the customer website or statistics. The drawback of the Mixpanel was the subscription based service model, compared to Google Analytics which was free. Mixpanel provides only 25 000 data points per month for free, if customer goes over that limit the data points become subject to charge. Mixpanel defines data point as an action in customer application or a request that is made to Mixpanel servers. For example one data point can be a click on a button or an application load. The amount of data points you spend depends completely on the structure of analytical system.

Luckily we were able to acquire Mixpanel partnership and amass 175 000 data points more in one month, which totaled to 200 000 free data points per month. That provided us with enough events to conduct planned A/B tests in our system.

5.1.2 Technical implementation and challenges

When working on a large scale enterprise system, it takes time to only understand the whole concept properly. When analytics and especially A/B testing is added to the equation, it makes the system even more complex.

Some may say that A/B testing is easy. Landing Page Optimization (LPO) can be actually quite straightforward, as there are multiple tools that will do almost everything automatically. You do not even need to know any syntax language like HTML or JavaScript to create your own A/B tests. Advanced tools will modify your page on the fly and will provide corresponding version of the page to the visiting user. In most cases there is no need to do any code changes to your site as it will be done by the used tool.

In some cases it is not possible to use tools that are intended for optimizing websites for your own project. In those situations own implementation is needed. You need to get your hands dirty and start implementing analytics in your own code. It is still not recommended to implement everything from scratch as there are many good services that will provide good interfaces and functionality to collect and analyze the gathered data. Though, it is still required to implement the logic for sending the data in the project.

Frequently people think that just implementing something like analytics is enough. They are wrong, like any project or service also analytics need maintenance. The requirements change, the code changes and behavior changes. If analytics are not updated accordingly the logic and funnels could grow old as well. What happened in the Kiosked was that at some point we were in such a rush that we were implementing new features and changing old requirements without up-keeping the analytics side of the project. This made us pay the price as later on we were receiving incorrect analytics and thus could not reliably say that the A/B test was based on well-established ground. What is important to remember here is that software engineers should always include the effort that comes from the analytics into the task estimations when planning sprints[8].

When including analytics in your own project, the integrity of the analytics should be also taken into account. User should not be able to abuse the analytics through multiple clicks for example on the kiosk buy button. Maybe one user with malicious intent could not distort or skew the results of the analytics drastically, but multiple users or a malicious program could easily do that.

5.2 A/B-testing results

Total of three A/B test were performed with Kiosked system. One test was run roughly for one month after the analysis was performed to determine statistical significance of the results. Not all of the results yielded significant results or even reached statistical significance.

5.2.1 First A/B test

The first A/B test was conducted in March 6th 2013 and was extended to 15th of April. We were testing two different versions of buttons: “buy now” vs “go to store”. The goal was to increase the

[8] Sprint is a timeboxed effort. It is one cycle of a fixed duration when software is being developed.

conversion of visitors to customers by engaging more users into the buying process. Version A was a button with “buy now” text and version B was button with “go to store” text. Three metrics were measured in our A/B test: hovers, which can also be considered as mouse overs on specific element on the computer screen, clicks and completion rate.

There were total of 880 hovers and 490 clicks. Below can be seen the distribution table between the two versions. Completion rate signifies the completion rate of the funnel – from mouse-hovering the media and hovering the “buy button” or “go to store” button respectively, and finally clicking it.

	Version A (buy now)	Version B (go to store)
Hovers	408	472
Clicks	228	262
Completion rate	13,51 %	15,52 %

Table 2: Results of the first A/B test

As can be seen from the Table 2 the completion rate for the Version B was a 2,01 percentage point higher than Version A. Furthermore based on the Pearson's chi-square statistical test with confidence level of 95% the hovers are meaningful. Unfortunately the click results are not meaningful with confidence level of 95%. The interesting aspect was that one week before ending our first A/B test, the hover results were not statistically significant either.

5.2.2 Second A/B test

The second test was started in July 7th 2013 and stopped in August 7th 2013. This time we were testing the conversion rate of traditional ”buy now” button versus ”buy now” button with a clickable link in the product title. When Kiosked launched the kiosks the only way to get to webshop was to use ”buy now” button. With the A/B test we wanted to test the concept with linking buy page to the product title as well as keeping ”buy now” button and see if it would bring us greater click-thru-rate. In version A we had only buy button leading to the store and in version B we included both: button and link.

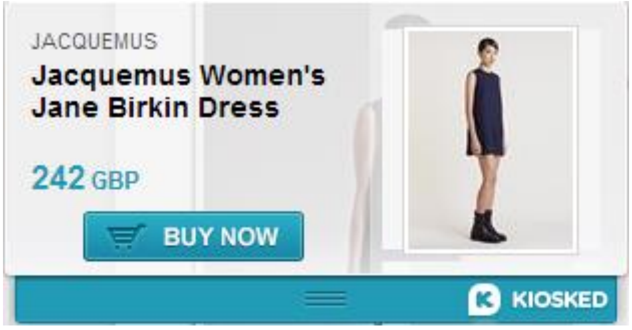


Figure 14: Kiosked A/B test with clickable title and buy button

The second test was not as successful as the first one. Total of 2804 buy button hovers and 742 clicks were recorded in that period of time. Unfortunately only measly 22 product title clicks were logged in the same period of time. However we did get quite many product title hovers as there were 984 total of them. We also logged clicks on the non-existent product title link in version A which yielded 6 clicks. Regrettably we lost our information regarding how many buy button clicks each version received respectfully during the testing period. This incident happened due poorly implemented solution and was not reviewed properly. On the other hand the measly amount of the 22 product link clicks related to the 742 button clicks would not mean any significant statistical difference.

	Version A (buy button only)	Version B (buy button with product link)	Total
Button hovers	-	-	2804
Button clicks	-	-	742
Product link hovers	-	984	-
Product link clicks	6	22	-

Table 3: The results of second A/B test

5.2.3 Third A/B test

The third test period was from August 21st 2013 to September 28th 2013. The plan was to test out a decision of product council that required kiosk to open automatically additional element when user hovered over kiosk. The element contained product information which can be seen from Figure 15. The previous default state of the kiosk view is showed in Figure 14. In Figure 14 user is required to click the bar at the bottom side of the kiosk to see additional product information.

We decided to test if the extra information flow would scare the users from interacting with the kiosk further and thus lowering the conversion rate of clicks on the buy button.

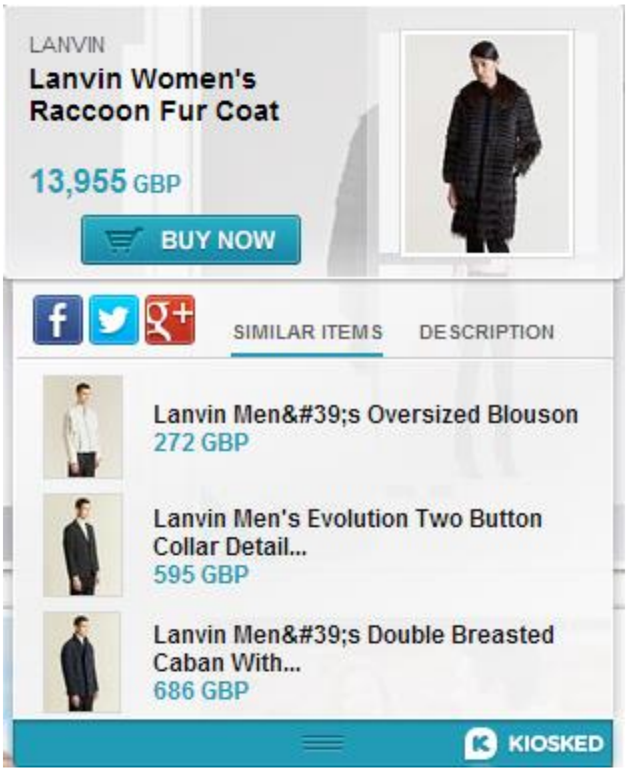

LANVIN
Lanvin Women's Raccoon Fur Coat
13,955 GBP
BUY NOW
SIMILAR ITEMS
DESCRIPTION
Lanvin Men&#39;s Oversized Blouson
272 GBP
Lanvin Men's Evolution Two Button Collar Detail...
595 GBP
Lanvin Men&#39;s Double Breasted Caban With...
686 GBP
KIOSKED

Figure 15: Open kiosk with additional product infromation

We only collected numbers of buy button clicks, with and without the feature enabled. In version A additional user interaction was required to obtain extra detail of the product while version B incorporated the automatic display of additional information.

	Version A (needed user interaction)	Version B (automatic display of additional information)
Clicks on buy button	354	221
Clicks on similar items	37	31
Clicks in total	391	252

Table 4: The results of the third A/B test

Based on the results and calculating the statistical significance with Person's chi-squared test we come to conclusion that the results are meaningful with over 95% confidence.

5.3 Results of conducted usability test

Usability test conducted by Charlotta was done on February 13th 2013 and it focused on usage of Kiosked service on mobile phones. Mobile phone usage among test group included: iPhone, Nokia E- and S-series and Nokia Lumia.

The objectives of the test were to find out answers to following questions:

- Do users notice the K-icon in the images?
- Do users find the kiosks and understand them?
- Do users understand how to navigate in the opened product list?
- Would users buy from the kiosks?

Most of the users stated that they noticed K-icon (11 users out of 12). One third tapped the icon without giving it a second though while rest did not interact with the K-icon before were told that there was something to interact with. Several users tried to tap the image instead and after

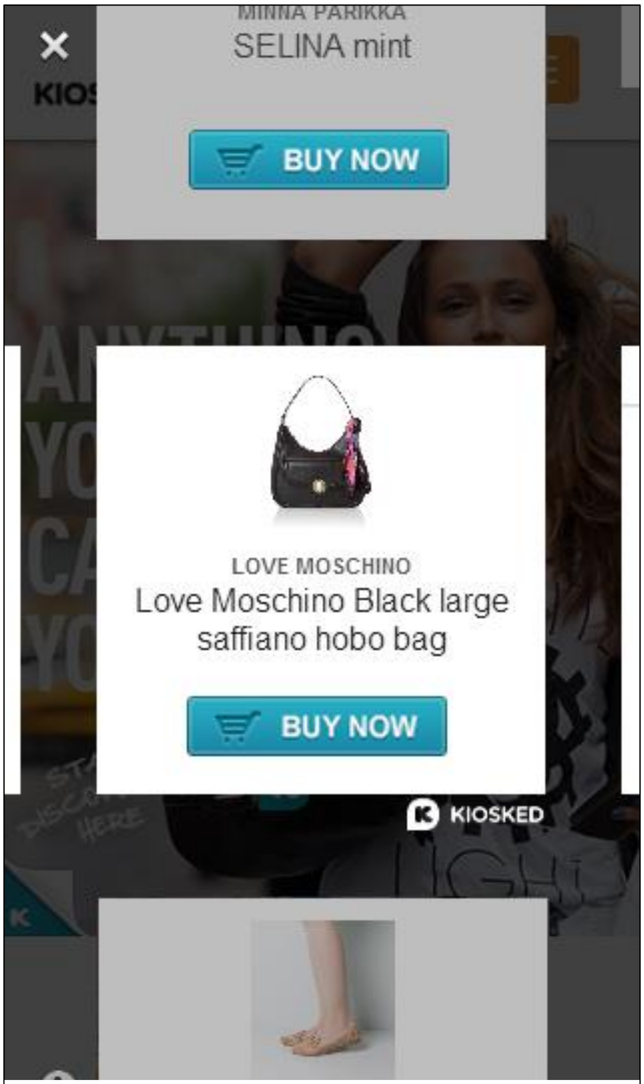


Figure 16: Mobile view of kiosks in image

noticing that nothing happens, they finally tapped K-icon.

Many of the users considered kiosks as a commercial but saw the purpose of the images that they worked as an interface for selling products. Almost all users (11/12) did not have any problems scrolling product list horizontally and vertically. Only one user was able to scroll horizontal only. He saw the additional products on top and bottom, but instead of scrolling tried to tap on them. One of the users commented on the scrolling: “Easy to use”.

Product relevancy was the most important aspect for the users. All users understood how to go about buying a product. Majority of the users questioned where the buy button would actually lead them. Quoting one of the user: “Not sure I would dare to click this”, when asked about the buy button and where the user think it would lead. On contrary another user commented: “Now I understand, this is good”, after clicking buy button and landing on the product provider site.

5.4 Comparison between usability test and A/B test results

Drawing the conclusion from the earlier results usability test and A/B test both drive the product or service into better user experience for customers. How this product superiority is achieved is completely different for both methods. Usability test is a more hands-on experience, where organizer can feel the pain and joy of the user. A/B test in turn approaches the problem through trial and error.

In our case we got surprisingly lucky as only one A/B test out of three was completely insignificant. As discussed in chapter 3.4.1 Strengths and weaknesses of A/B testing that many A/B tests will actually fail. A/B testing is not a Holy Grail that will find and fix all the problems by itself. Sometimes as many as close to ten A/B test can be insignificant before getting meaningful results (Chopra, 2011).

Both tests require different competence and skills. Usability test is more of a qualitative test while A/B test is a quantitative and requires more technical person to actually implement the logic into the software code when own system is in question. As mentioned earlier, there are a lot of Landing Page Optimization (LPO) articles and tools to improve a site or a store of the company. The case is not that easy with enterprise based solution, where the system is much more than just a website on the internet. On the other hand usability test requires more preparation time for each test than A/B test in comparison, assuming that the fundamental of A/B testing is in place. Additionally analysis of the data collected from the usability tests requires more time. Transcription of the audio or video recordings is a time consuming as well as anesthetic. Taking notes while conducting usability test may make supervisor to unintentionally overlook some of the actions and comments from a user.

Results that can be gain from the usability test can be actually analyzed more carefully and even split easier into a user segments. For example if evaluator will notice that certain type of people prefer particular approach in using the product or service they can communicate that to product owners. A/B test in turn provides general overview of the results where all the user actions are mixed and the bigger segment of the users will win.

When doing usability tests one should not forget the evaluator effect. Results gathered form the usability test by two different evaluators can be varying (Hertzum & Jacobsen, 2001). With A/B testing there is also an error margin for implementation and collection of the data as well, but one does not need to be a good evaluator to implement a basic A/B test.

What kind of method one should use then, usability testing or A/B testing? When two designs exist with different focuses or goals A/B testing can be invaluable tool. Usability testing in comparison provides better tool for actually understanding users and their thoughts. Usability test answers more of a question *why*, while A/B test answers question *which* or *how many*.

Usability testing	A/B testing
Why users are not clicking the link?	Which link generates more clicks?
Why users do prefer that specific pricing model?	Which pricing model is performing best?
Why people are not interacting with the site?	Which website layout generates most sign ups?
Why I have so low amount of visitors?	Which email campaign performs better

Table 5: Usability testing vs A/B testing, question comparison

6 Discussion

The examination of this thesis was targeted on the A/B testing and usability assessment methods, a topic that is reasonably new and has not been yet widely studied. The first research question emphasizes on the strengths and weaknesses of both methods as well as proposes how both methods can be combined. The second research question highlights how A/B tests could be realized in small companies and what could be taken into consideration when conducting A/B tests.

6.1 First research question

Both approaches A/B testing and usability assessment methods are crucial in understanding the customer. What we learned during the testing period on this subject is that usability assessment methods are great tools to finding out the most critical issues in your system. When the issue has been identified and the solution has been proposed, A/B test can help in asserting the proposal with real case scenario.

Usability assessment methods and especially usability testing can provide behavioral insight of users. In turn, A/B test can measure very small performance difference with high statistical significance and can resolve trade-offs between conflicting guidelines. Based on the research and empirical results usability assessment methods were found to be a great first stage practices evaluating a product. A/B test can be used as a follow up methods which will help to fine-tune particular aspects and segments of the product.

Time effort was not measured when conducting usability test and A/B tests. Reason for that was uncertainty whether or not there would be any testing after first usability test. Only one usability test was done at that time and no clear plan for A/B was made. Measuring implementation time of the A/B test and the effort used to implement different variations compared to the usability assessment methods practices is important aspect of comparison and should be studied further.

Breaking down the resource requirements to prepare usability test or A/B test should be further explored. A/B test can be thought as an architectural based technique. Depending on the code quality of the project, implementing A/B tests can either be easy or difficult.

6.2 Second research question

We learned a lot by only performing three A/B tests in our company. When conducting A/B tests preparation and understanding of the system is a key. In a rush many things can be overlooked, and even small mistake can ruin the whole A/B test. One single misplaced request can make data worthless. Similar issue happened with our second A/B test where we should have separated buy button clicks with enabled and disabled title link. Thence the test lost its importance as we could not calculate whether the product link actually affected the amount of clicks on the buy button.

Additionally our implemented A/B tests solution lacked persistence. The solution did make the split between versions based on a user IP address, but did not save that option on the user computer for example in a cookie. Without the cookie a user who visited a kiosked image from office network could see different version from home network due to different IP addresses. Improvement could be made so, that first system would check whether or not cookie with selected option is already saved. If so, the saved option would be displayed. Otherwise, a version to show would be chosen based on the IP calculations and the generated option would be then saved into cookie. This way A/B tests would not confuse users much as they would see only one version regardless of their location. Surely this would not solve the issue when the user would switch a computer or would clear their browser cache or cookies. Today people tend to consider computer and especially smartphone as personal devices meaning it is more likely that user will be using only his or her personal devices to access services.

To further increase the relevance of the A/B test, it is important to develop deep conversion funnel. In ideal solution conversion funnel should start from the first user interaction and end at desired goal. In case of Kiosked that would mean:

1. User landing on a website that has a kiosked media
2. User noticing media (mouse over the media)
3. User exploring products or services in a media (clicking products and finding information in a kiosk)
4. User engaging the media (clicking buy button inside kiosk)
5. User reaching the goal (buying the product from the third party webshop)

In most cases that is difficult to achieve, particularly when there are third parties in cooperation. Creating a complete funnel on your own side is hard enough, but then additional data and information is needed from the cooperating company.

We believe that whether you are developer, manager or advertiser it does not matter, because everyone should understand at least the basics of usability and its importance. If your manager understands the practice and benefit of the A/B test and usability assessment methods the implementation of both methods can get a great start and it is possible to improve your business in due time. Point is that the most challenging aspect of usability and A/B testing can be convincing your supervisor or manager to implement and start with the practices themselves. The smaller the organization the more precious the company is to the Chief executive officer (CEO). It is the CEO who wants to be responsible on how the things are done and how they look in the company. This is where A/B testing can prove valuable. All the proposals and ideas made by management related to design, usability or functionality can be tested with A/B testing. People usually have their own unique perspectives and biases when dealing with web pages, design and functionality. The most important lesson that should be taken from this work is that all assumptions must be left aside to be able to understand and improve your business. For the successful business there is room only for the knowledge.

The bigger companies have their own challenges regarding the improvement to usability and conversion rate. Without going into much detail, usually the bigger the company the more difficult it is to make initiative and drive the idea forward. There have been quite a lot of studies of companies and structural inertia. Sometimes big companies have some kind usability assessment practice already in place, which makes it tremendously easier to take off.

The benefits of the large companies compared to small ones, are huge resources, such as money and people. Sometimes small to medium-sized companies are reluctant to invest into usability tests or conversion rate optimizations just because the resources are needed somewhere else. As mentioned before, the A/B testing does not even require a great amount of resources, but even that small step can sometimes seem to be insurmountable. Often many tasks or activities are seen more important than understanding your users, even if knowing your users will help the company in the long run. We stumbled upon many situations where some of the features were in development priority list even though there was no clear evidence whether users actually needed them or used them. Strong opinions of the management mattered the most. Achieving a product with great usability is a process, not an opinion of a one person.

7 Conclusion

Small companies prefer to practice conversion optimization with A/B testing rather than conducting traditional usability testing or introduce usability inspection methods. Usability assessment methods do not require great amount of candidates. In usability testing only five test users can point out most of the usability issues in a system. Issues that are brought up during usability assessment methods should be validated with A/B testing. Even though usability assessment methods can point out most of the problems, it is still only a laboratory-like experiment or a recommendation from an expert. Context in real case could be slightly different. Both methods are needed to achieve complete understanding of the users.

A/B testing is used mostly in conversion optimization. One may think that it does not relate to usability but it is still associated with human behavior and preferences. Finding a starting point for A/B test can be difficult. Which element or part of the system you should A/B test? In both methods, usability assessment methods and A/B testing, the answer lies in a goal. What is expected from the user and what user should achieve? A/B test tends to answer to question *which* or *how many* while usability assessment methods and especially usability test answers to question *why*. Usability assessment methods work well as an initial evaluation while A/B test can be used as a follow up method to fine-tune product or service even further.

Collected feedback from usability assessment methods can prove valuable. There is no easier way to start an A/B test than with validating the ideas or suggestions collected in a feedback. Also ideas that come from a management can be tested with A/B testing. A/B testing requires definitely more studies and research. Moreover the difference and importance of both approaches should be explored further.

Challenging part in practicing usability assessment methods or conducting A/B test is convincing management to actually accept the fact that the process and methods are needed. Company can not claim that they are user-centered or user driver company if they do not actually engage with users. Usability assessment methods and A/B testing is not only about making a product look good or improving some conversion rates – it is about making the product successful.

References

Benjamini, Y. & Yekutieli, D., 2001. The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics*, 29(4), pp. 1165-1188.

Benyon, D., Turner, P. & Turner, S., 2005. *Designing Interactive Systems: People, Activities, Contexts, Technologies.* Essex: Pearson Education UK.

Bock, G.-W., Kuan, H. H. & Vathanophas, V., 2005. Comparing the Effects of Usability on Customer Conversion and Retention at E-Commerce Websites. *Hawaii International Conference,* Issue 38, p. 174.

Burk, S., 2006. A Better Statistical Method for A/B Testing in Marketing Campaigns. *Marketing Bulletin,* 3(17), pp. 1-8.

Chopra, P., 2011. *Appsumo reveals its A/B testing secret: only 1 out of 8 tests produce results.* [Online]
Available at: http://visualwebsiteoptimizer.com/split-testing-blog/a-b-testing-tips/
[Accessed 15 October 2013].

Cohen, J., 2009. *Easy statistics for AdWords A/B testing, and hamsters.* [Online]
Available at: http://blog.asmartbear.com/easy-statistics-for-adwords-ab-testing-and-hamsters.html
[Accessed 3 June 2013].

Conversion Rate Experts, n.d. *Split-testing 101: A quick-start guide to conversion rate optimization.* [Online]
Available at: http://www.conversion-rate-experts.com/cro-tips/
[Accessed 14 April 2014].

Dumas, J. S. & Salzman, M. C., 2006. Usability Assessment Methods. *Review of Human Factors and Ergonomics*, 1 April, pp. 109-136.

Eisenberg, B., 2008. *Hidden Secrets of the Amazon Shopping Cart.* [Online]
Available at: http://www.grokdotcom.com/
[Accessed 15 May 2013].

Farakh, M. A., 2013. *Most of your AB-tests will fail.* [Online]
Available at: http://www.jitbit.com/news/185-most-of-your-abtests-will-fail/
[Accessed 15 September 2013].

Foss, N. J., 1997. *Resources, Firms, and Strategies: A Reader in the Resource-based Perspective.* New York: Oxford University Press.

Gerken, J., Bak, P., Jetter, HC., Klinkhammer, D. & Reiterer, H., 2008. *How to use interaction logs effectively for usability evaluation.* Konstanz: Bibliothek der Universität Konstanz.

Gofman, A., 2007. Consumer driven multivariate landing page optimization: overview, issues and outlook. *The IPSI BgD Transactions on Internet Research 3.2,* pp. 7-9.

Google, 2013. *Google Analytics.* [Online]
Available at: http://www.google.com/analytics/
[Accessed 9 September 2013].

Gray, W. & Salzman, M., 1998. Damaged merchandise? A review of experiments that compare usability. *Human-Computer Interaction,* 13(3), pp. 203-261.

Hancock, P., Pepe, A. & Murphy, L., 2005. Hedonomics: The power of positive and pleasurable ergonomics. *Ergonomics in Design,* pp. 8-14.

Hertzum, M. & Jacobsen, N. E., 2001. The Evaluator Effect: A Chilling Fact About Usability Evaluation Methods. *Publishing models and article dates explained,* 13(4), pp. 421-443.

Hollingsed, T. & Novick, D., 2007. *Usability Inspection Methods after 15 Years of Research and Practice.* New York, ACM, pp. 249-255.

Holzinger, A., 2005. Usability engineering methods for software developers. *Communications of the ACM - Interaction design and children,* 48(1), pp. 71-74.

International Organization for Standardization, 1998. *ISO 9241-11.* s.l.:s.n.

Jeffries, R. & Desurvire, H., 1992. Usability testing vs. heuristic evaluation: was there a contest?. *ACM New York,* 24(4), pp. 39-41.

Kanich, C., Kreibich, C., Levchenko, K., Enright, B., Voelker, GM., Paxson, V. & Savage, S., 2008. *Spamalytics: An Empirical Analysis of Spam Marketing Conversion.* New York, ACM, pp. 3-14.

Kiosked, 2013. *Kiosked.* [Online]
Available at: http://www.kiosked.com
[Accessed 10 August 2013].

Lane, D. M., 2013. *HyperStat Online, Statistics Solutions.* [Online]
Available at: http://davidmlane.com/hyperstat/A42408.html
[Accessed 1 October 2013].

Lofgren, L., 2013. *Quora: What are the best A/B testing tools for SaaS products.* [Online]
Available at: http://www.quora.com/A-B-Testing/What-are-the-best-A-B-testing-tools-for-SaaS-products
[Accessed 1 October 2013].

Mesibov, M., 2011. *The Right Test at the Right Time.* [Online]
Available at: http://blog.abovethefolddesign.com/2011/09/08/the-right-test-at-the-right-time
[Accessed 20 April 2014].

Muller, M. J., Matheson, L., Page, C. & Gallup, R., 1998. Methods & tools: participatory heuristic evaluation. *ACM,* 5(5), pp. 13-18.

Nassimian, M. B., 2011. *A/B Testing SEO friendly.* [Online]
Available at: http://online-behavior.com/testing/ab-testing-for-seo
[Accessed 15 February 2014].

Nielsen, J., 1990. Heuristic evaluation of user interfaces. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 249-256.

Nielsen, J., 1993. *Usability Engineering.* California: Academic Press.

Nielsen, J., 1994. *Usability Inspection Methods.* New York, ACM, pp. 413-414.

Nielsen, J., 1995. *Nielsen Norman Group.* [Online]
Available at: http://www.nngroup.com/articles/ten-usability-heuristics/
[Accessed 5 1 2014].

Nielsen, J., 2000. *Nielsen Norman Group.* [Online]
Available at: http://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/
[Accessed 19 January 2014].

Nielsen, J., 2005. *Putting A/B Testing in Its Place.* [Online]
Available at: http://www.nngroup.com/articles/putting-ab-testing-in-its-place/
[Accessed 20 April 2014].

Nielsen, J., 2012. *Nielsen Norman Group.* [Online]
Available at: http://www.nngroup.com/articles/thinking-aloud-the-1-usability-tool/
[Accessed 19 January 2014].

Nieminen, M., 2013. *User Experience lecture.* Espoo: Aalto University.

NIST/SEMATECH, 2012. *e-Handbook of Statistical Methods.* [Online]
Available at: http://www.itl.nist.gov/div898/handbook/eda/section3/eda3674.htm
[Accessed 6 June 2013].

Plackett, R. L., 1983. Karl Pearson and the chi-squared test. *International Statistical Review/Revue Internationale de Statistique,* 51(1), pp. 59-72.

Polson, P., Lewis, C., Rieman, J. & Wharton, C., 1991. Cognitive walkthroughs: a method for theory-based evaluation of user interfaces. *Proceedings of the Conference on Human,* 36(5), pp. 741-773.

Quora, 2012. *Quora: What are the best A/B and multi variant testing tools available?.* [Online]
Available at: http://www.quora.com/A-B-Testing/What-are-the-best-A-B-and-multi-variant-testing-tools-available
[Accessed 1 October 2013].

Roscoe, J. T. & Byars, J. A., 1971. An Investigation of the Restraints with Respect to Sample Size Commonly Imposed on the Use of the Chi-Square Statistic. *Journal of the American Statistical Association*, 66(336), pp. 755-759.

Rubin, J. & Chisnell, D., 2008. *Handbook of Usability Testing: Howto Plan, Design, and Conduct Effective Tests.* Indianapolis, IN: Wiley Publishing, Inc.

Salmoni, A. J. & Gupta, S., 2012. *Quora - How long should you run an A B test on your site before you declare one a winner.* [Online]
Available at: http://www.quora.com/A-B-Testing/How-long-should-you-run-an-A-B-test-on-your-site-before-you-declare-one-a-winner
[Accessed 15 September 2013].

Shermer, M., 2002. *The Skeptic Encyclopedia of Pseudoscience 2 volume set.* 1st ed. California: ABC-CLIO Inc..

Siroker, D. & Koomen, P., 2013. A/B Testing: The Most Powerful Way to Turn Clicks Into Customers. In: *A/B Testing: The Most Powerful Way to Turn Clicks Into Customers.* New Jersey: John Wiley & Sons, pp. i-v.

Steiner, I., 2008. *eCommerce Bytes.* [Online]
Available at: http://www.ecommercebytes.com/cab/abn/y08/m02/i22/s02
[Accessed 11 July 2013].

Sumner, G., n.d. *A/B and Multivariate testing.* [Online]
Available at: http://www.sitefinity.com/resources/whitepapers/download/discover-the-most-effective-version-of-your-website-with-a-b-and-multivariate-testing
[Accessed 2 February 2014].

Wharton, C., Rieman, J., Lewis, C. & Polson, P., 1994. *The Cognitive Walkthrough Method: A Practitioner's Guide.* New York: John Wiley & Sons, Inc.

Visual Website Optimizer, n.d. *Case Studies & Success stories.* [Online]
Available at: http://visualwebsiteoptimizer.com/case-studies.php
[Accessed 15 May 2013].

Visual Website Optimizer, n.d. *Customer List.* [Online]
Available at: http://visualwebsiteoptimizer.com/customers.php
[Accessed 15 May 2013].