



DEGREE PROJECT IN COMPUTER ENGINEERING, FIRST CYCLE
STOCKHOLM, SWEDEN 2017

Conversion Rate Optimization of E-Commerce using Web Analytics and Human-computer Interaction Principles

*An in-depth Quantitative Approach to
Optimization of Conversion Rates*

UTSAV KAUSHIK and ANTONIO GRONDOWSKI

Conversion Rate Optimization of E-Commerce using Web Analytics and Human-computer Interaction Principles

An in-depth Quantitative Approach to Optimization of Conversion Rates

Utsav Kaushik and Antonio Grondowski

2017-05-12

Bachelor's Thesis

Examiner

Gerald Q. Maguire Jr.

Academic adviser

Anders Västberg

Abstract

For an e-commerce business to grow, there are many ways one could try to improve the business in order to gain greater reach and increase sales. One of the main goals of such businesses is to convert as many visitors as possible into customers. Even though many e-commerce businesses already have web analytics tools installed, e-merchants find difficulty in identifying *where* to start optimizing, *what* data to extract from analysis reports, and *how* to make use of such data in order to produce a successful design that will increase the conversion rate. The purpose of this thesis is to (without spending resources on marketing-related factors) guide companies to find a low cost and efficient way to increase the conversion rate by creating well-thought-through designs based on analytic data, qualitative research, and human-computer interaction principles.

Google Analytics, a web analytics tool, was used in identifying high-valued pages to optimize and to identify demographics/target groups, while qualitative e-commerce related research was used to shape design-proposal hypotheses. This, along with two A/B tests conducted using Optimizely, is the basis for the guidelines and conclusions.

The results of both A/B tests showed an increase in conversions with designs highlighting: evidence of a secure shopping environment, incentives that will attract visitors to buy, and by removing auxiliary navigation elements at the check-out page. The evaluation of the results and its statistical significance was done using both Optimizely's statistical engine and null hypothesis testing. The increases in conversions were not statistically significant per Optimizely; however, they were significant using traditional statistics.

In conclusion, using metrics such as high exit-rates combined with many page views and high revenue-generating pages will allow e-merchants to identify where to start their optimization process. Furthermore, to know what valuable data needs to be extracted, one should seek the data that needs to be inserted into HCI concepts, such as personas and scenarios. This, along with qualitative research allows designers to create well-thought out design-proposals that will potentially lead to an increased conversion rate.

Keywords

Conversion rate optimization, A/B testing, E-commerce, Quantitative Research, Interaction Design, Human-computer interaction, HCI, Web Design, Statistical Inference, Web Analytics.

Sammanfattning

För att få en e-handelsbutik att växa finns det många arbetsområden man kan försöka förbättra för att nå ut till fler samt öka försäljning. Ett av huvudmålen för dessa butiker är att konvertera så många besökare till kunder som möjligt på sin hemsida. Även om många e-handelsbutiker redan har webbanalytiska redskap till sitt förfogande, har många tjänsteleverantörer svårigheter med att fastställa *var* på hemsidan det skall optimeras, *vilken* data som ska hämtas från analysrapporter, och *hur* man använder sig av dessa data för att skapa en lyckad design som kommer öka konverteringsgraden. Syftet med avhandlingen är att, utan marknadsföringsrelaterade investeringar, vägleda företag till billiga och effektiva sätt att öka konverteringsgraden. Detta ska uppfyllas genom att skapa väl genomtänkta designers grundade på analytisk data, kvalitativ forskning, samt människa-datorinteraktions principer.

Webbanalysverktyget Google Analytics användes för att identifiera högt värderade sidor att optimera och demografier/målgrupper medan kvalitativ e-handels-relaterad forskning användes för att forma hypoteser kring designförslagen. Detta, tillsammans med två A/B tester som genomfördes med hjälp av Optimizely, är grunden till riktlinjerna och slutsatserna.

Resultaten från båda testerna visade en ökning i konverteringar med designers som framhäver; övertygande eller bevis för en säker handelsmiljö, incitament som kommer locka besökare att handla, och genom att ta bort extra navigeringselement vid kassasidan. Utvärdering av resultaten och dess statistiska signifikans gjordes med Optimizelys statistiska motor såväl som egen nollhypotes prövning. Ökningarna av konverteringar var inte statistiskt signifikanta enligt kalkyl från Optimizely, men lyckades nå signifikans enligt traditionell statistik.

Sammanfattningsvis, med hjälp av mätvärden så som höga utgångsfrekvenser i kombination med högt antal sidvisningar samt höga intäktsgenererande sidor, kan tjänsteleverantörer nu identifiera var man kan påbörja optimeringsprocessen. För att veta vilken värdefull data man bör extrahera skall man ta reda på vilken data som behövs för att stoppa in i Människa-datorinteraktion (MDI) koncept, som personans och scenarier. Detta, tillsammans med kvalitativ forskning, tillåter webbdesigners att skapa väl genomtänkta designförslag som förhoppningsvis leder till en ökad konverteringsgrad.

Nyckelord

Konverteringsoptimering, A/B-testning, e-handel, kvantitativ studie, interaktionsdesign, människa-datorinteraktion, MDI, webbdesign, statistisk inferens, webbanalys.

Acknowledgments

We would like to thank Maria Richardsson for giving us the opportunity to do our project on Nordic Design Collective's website. We also want to thank Kristoffer Richardsson for not only helping us obtaining as much information as possible about Nordic Design Collective's website, but also for continuously pushing us and helping us being critical throughout the project. Finally, a great thanks to Gerald Q. Maguire Jr. for his guidance in shaping this thesis and his feedback concerning what to include and what was unclear.

Stockholm, May 2017

Utsav Kaushik and Antonio Grondowski

Table of contents

Abstract	i
Keywords	i
Sammanfattning	iii
Nyckelord	iii
Acknowledgments	v
Table of contents	vii
List of Figures	xi
List of Tables	xiii
List of acronyms and abbreviations	xv
1 Introduction	1
1.1 Background	2
1.2 Problem definition	2
1.3 Content	3
1.4 Purpose	3
1.5 Goals	3
1.6 Research Methodology	3
1.7 Delimitations	4
1.8 Structure of the thesis	4
2 Background	5
2.1 Conversion	5
2.2 Working with Conversion Optimization	5
2.3 Getting to Know Your Customers: Developing Personas	6
2.4 Scenarios	7
2.5 Exit rate and using it to determine where to start to optimize	7
2.6 Acceptable Exit Rates for a page	8
2.7 Designing for Interaction - Laws and Principles	8
2.8 Important features of an e-commerce store	13
2.9 Strengths and Weaknesses of Physical and E-commerce Stores	14
2.10 Consumers and Companies' view of them	14
2.10.1 Clear Information about Products	15
2.10.2 Total price	15
2.10.3 Good search function.....	16
2.10.4 Easy to navigate	16
2.10.5 Clear Contact- and Customer Service Information	16
2.10.6 Registration should be optional.....	16
2.11 Differences in Male and Female Consumer Behavior (in e-commerce)	17
2.12 Sample Data	17
2.13 Google Analytics	17
2.13.1 Visitor Segmentation.....	18
2.13.2 Sales Performance	18
2.13.3 Sessions	18

2.14	Related work	18
2.14.1	Background on conversion optimization	18
2.14.2	Importance of knowledge about target groups	19
2.14.3	Design areas affecting conversion rate	19
2.15	Summary	19
3	Methodology	21
3.1	Research Process	21
3.2	Research Paradigm	22
3.3	Data Collection	23
3.3.1	Google Analytics	23
3.3.2	Optimizely	23
3.3.3	Sampling	24
3.3.4	Sample Size	24
3.3.5	Demographics and Target Groups	24
3.4	Experimental design/Planned Measurements	24
3.4.1	Qualitative versus Quantitative Research	25
3.4.2	A/B Testing vs Multivariate Testing vs Multi-Armed Bandit Testing	25
3.4.3	A/B Testing	26
3.4.4	Benefits and limitations of A/B-tests	27
3.4.5	Test environment/test bed/model	28
3.4.6	Reliability	29
3.4.7	Statistical Significance	30
3.4.8	Validity using Sequential Testing combined with FDR control	31
3.5	Planned Data Analysis	33
3.5.1	Data Analysis Technique	33
3.5.2	Software Tools	33
3.6	Evaluation framework	34
4	Results	35
4.1	Extracted Google Analytics User Data for Personas	35
4.1.1	Step 1: Audience → Demographics → Overview	35
4.1.2	Step 2: Audience → Demographics → Gender → Secondary dimension = Age	36
4.1.3	Step 3: Audience → Interests → Overview → Set Segment: Males, Segment: Females	36
4.1.4	Step 4: Audience → Mobile → Devices	37
4.1.5	Step 5: Customization → Hourly & Daily Engagement (Hour, Date & Day)	38
4.1.6	Step 6: Audience → Geo → Location	39
4.2	Persona One – Karin	39
4.2.1	Scenario - Karin	40
4.2.2	Scenario Insights - Karin	40
4.2.3	Persona Two – Stefan	40
4.2.4	Scenario - Stefan	41
4.2.5	Scenario Insights - Stefan	41

4.3	Finding where to Start - Inspect highly valued and trafficked pages	42
4.4	A/B testing Design Proposal 1 – Before and after	42
4.5	Results of A/B test 1: Info-box and plus and minus functions	46
4.6	Going Where the Rainbow Ends - Inspect the Check-out Funnel Page(s).....	46
4.7	Results of A/B test 2: Removing Check-out Navigation.....	51
5	Analysis	53
5.1	Major results	53
5.1.1	Interpreting Optimizely’s Statistical Engine to make business decisions.....	53
5.1.2	A/B Test 1	53
5.1.3	A/B Test 2	54
5.2	Reliability Analysis.....	54
5.3	Validity Analysis	57
5.4	Discussion	58
5.4.1	Challenges due to Poor Integration of Software Tools with One Another	58
5.4.2	Time to reach significance	59
5.4.3	Ambiguity in Stats Engine	60
6	Conclusions and Future work	61
6.1	Conclusions	61
6.2	Limitations	63
6.3	Future work.....	64
6.4	Required reflections	64
	References	67
	Appendix A: Statistical Significance Calculation Using the Null Hypothesis.	71
	Appendix B: Detailed results	73

List of Figures

Figure 2-1:	Survey answer distribution among consumers and companies to the question: How Important are the following factors when shopping from an online store? [1p. 43]	15
Figure 3-1:	Difference of increase in Type 1 Error rates between traditional and sequential testing when continuously monitoring (4 times) [34].	32
Figure 3-2:	Multiple testing problem	33
Figure 4-1:	Session distribution among males and females	35
Figure 4-2:	Percentage of sessions of various target groups	36
Figure 4-3:	An overview of male and female interests	37
Figure 4-4:	Within mobile device usages, the percentage of sessions made through specific devices	38
Figure 4-5:	Percentage of sessions hour by hour (hourly engagement on the website).	38
Figure 4-6:	Distribution of sessions amongst countries	39
Figure 4-7:	Persona of the main target group, Karin, 31.	40
Figure 4-8:	Persona of the main target group – Stefan, 28	41
Figure 4-9:	Original design of a product page	43
Figure 4-10:	Experimental design of a product page	44
Figure 4-11:	A/B test results of Design 1: Info-box and + and - function ...	46
Figure 4-12:	Funnel visualization of sessions flowing from the cart page to completing a purchase.	48
Figure 4-13:	The original version of the check-out page.	49
Figure 4-14:	The experimental design variation of the check-out page.	50
Figure 4-15:	A/B test results of Design 2: Removing Check-out Navigation.	52
Figure 6-1:	Decision Tree for A/B test culmination	62

List of Tables

Table 2-1:	Principles [3, pp.129-138].....	10
Table 2-2:	Important features regarding website's layout, according to customers [1, p. 40].	13
Table 2-3:	“Why I purchased in a physical store instead of an online store” [1, pp. 42–43]	14
Table 2-4:	“Why I purchased in an online store instead of a physical store” [1, pp. 42–43]	14
Table 3-1:	Benefits and limitations of A/B-tests [29]	28
Table 4-1:	High-valued pages	42
Table 5-1:	Key z-statistics values for A/B test one and two	58

List of acronyms and abbreviations

CRO	Conversion rate optimization
E-commerce	Electronic commerce
FDR	False discovery rate
GfK	Gesellschaft für Konsumforschung
HCI	Human—computer interaction
KPI	Key performance indicator
URL	Uniform resource locator
WWW	World wide web

1 Introduction

This chapter initially presents the context for this thesis project, and then continues to describe the problem, necessary background knowledge, purpose, and the goal of the thesis project. Thereafter, methods will be described, the chosen delimitations, and the general structure of this thesis.

Continuous technological modernization has fostered increased electronic commerce (e-commerce). In 2015, Swedish e-commerce had a turnover of 50.1 million SEK, which corresponds to 6.9% of the total retail sales and was a 19% growth from the previous year [1p. 6].

Although the goal of every retail interaction, whether physical or online, is to convert visitors into paying customers, the (psychological) science of shopping differs between the physical and online experience [2]. According to P. Underhill, a retail anthropologist, some strategies used to convert visitors into paying customers in a physical store entail [2]:

- Building a store with a route that forces customers to go counter-clockwise, as this will make individuals more prone to buy more since the majority of people are right-handed, thus they have a hand free to put products into their cart.
- Placing product such as perfume and cosmetics at the entrance of a store will induce self-awareness; hence visitors will let their guards down.
- Design and aesthetics of some grocery stores are deliberately made to look cheap by having simple lighting and exhibiting products in cartons to create an illusion of cheapness.
- Product placement of the most essential foods, such as bread, milk, and vegetables is made specifically to put them furthest away from each other to make the path long enough for impulse purchases.

In online retail stores (e-commerce) the ability to touch products and have face-to-face communication with sales-people are non-existent leading to a digital sales strategy. In such a strategy, factors such as the web page's layout and the ease of completing a transaction become crucial.

Digital sales strategies are essential for e-commerce companies to remain competitive [1p. 5]. As e-commerce companies continuously improve their website design and communication to increase sales, the organization's designers often use their personal knowledge and knowledge of the organization in their design [3p. 43]. However, basing a design on personal taste in a heavily trafficked website is risky, hence statistical data is needed to support deciding which design is most effective and produces greater sales. Retailers that lack a well thought-out digital strategy will be overtaken in the marketplace [1p. 5]. Key questions thus become: How does one know where to start optimizing? Where does one find relevant user-behavior data and how can one make sense of this data?

One goal an e-commerce company has when people enter their website is to retain the visitor on the site as long as possible. This is done because as long as a visitor is active, the chances for additional purchases increase. If there is a specific

location on the site where visitors tend to drop off, with a so-called higher “exit-rate” than other parts of the site; then this specific location is worth evaluating and adjusting. Another outcome that an e-commerce company wants is for their visitors to purchase products, thus become paying customers. The proportion of those who become a paying customer is called the “conversion rate”. The overall goal of every e-commerce website is to decrease the exit-rate and increase the conversion rate, thus this thesis intends to answer the following questions:

- How to evaluate e-commerce data and make a design that leads to an increase in conversion rate and a decrease in exit rates?
- Which factors affect visitors during the buying process and how can these factors be exploited to increase the conversion rate?

1.1 Background

The continuous growth of e-commerce market and the impact e-commerce has on the market makes it vital for e-commerce stores to test how to maximize profits in order to give the business advantages in relation to their competitors.

In order for e-commerce companies to understand how to build their web page, the key ingredient is knowledge of how potential and current customers behave within the company’s website. From a conversion viewpoint, it is vital to identify the factors that have the greatest significance upon customers staying on the website. Web analytic data collection tools allow us to retrieve data about the user audience and their behavior can be exploited in combination with human-computer interaction (HCI) theories and design principles to define a customized way to improve a website’s design, rather than use general principles that may not work for the website’s specific demographics.

A common mistake designers make is to try to improve everything on the web page at once, rather than making iterative changes only where these changes actually matter [4]. It is often the case, that it is difficult to judge what actually affects the company, whether it is an experimental result or not. These experiments are typically conducted using A/B testing, in which two variations of the same design compete against each other with 50% of visitors using version A (the control) and 50% using version B (the variation). The company can then make a firm decision on which design is best based on the results of this testing [5].

1.2 Problem definition

A big challenge within conversion optimization is to identify where to start the optimization process. The challenge grows even greater for websites with a vast number of pages, such as an e-commerce store. Our main questions are:

- How does one identify the most significant pages to optimize within a specific website, with respect to conversion rate? Having identified these pages, how can the website be optimized to increase the conversion rate?

- Another challenge is to extract the most appropriate user data which both reflects the website's target group(s) and can be used in future design proposals.

The final question above leads to our subquery:

How does one identify the most valuable web analytic data? Using this data, how can one make a successful design based on human-computer interaction (HCI) design theories and principles?

1.3 Content

The project was conducted in co-operation with Nordic Design Collective AB (<http://www.nordicdesigncollective.se>), an e-commerce home decor & furniture store selling paintings, posters, minor furniture, etc. The company's ambition is to help new and independent Nordic designers to sell their products.

1.4 Purpose

The purpose of the project is to increase the conversion rate (and possibly decrease the exit-rate) by improving the website's design by considering HCI theories and principles and based on the analysis of the web traffic to this site. More specifically, the purpose is to identify the weaknesses and strengths of Nordic Design Collective AB's current web page. Nordic Design Collective could benefit monetarily and its user base will benefit from an improved design (i.e., on that provides ease-of-use and an intuitive user experience).

1.5 Goals

The goal of the project is to present design suggestions based on analysis of web traffic and concretely test these suggestions to see whether they produce a more successful e-commerce website by enhancing the user's experience and achieving a higher conversion rate. This will hopefully create guidelines of how to improve an e-commerce website's design and communications by using tools such as Google Analytics (see Section 2.13) and Optimizely (see Section 3.3.2).

1.6 Research Methodology

The work began with a literature study to collect relevant knowledge necessary to proceed with the rest of this thesis project. Subsequently, quantitative research was conducted in the form of A/B tests in which statistical data is used to decide whether a design was successful or not. This method has very clear outcomes; hence one can draw conclusions with relative ease. A/B testing is the best modern approach for testing two competing designs. In contrast, other approaches, such as multivariate-testing and multi-armed bandit test, were not considered since when using the results of these other types of testing it would be more difficult to judge what aspects of the design affected the conversion rate.

1.7 Delimitations

We will not focus on any marketing-related factors that increase the flow of visitors to the site, such as clearance sales, newsletters, etc. Our focus is to examine conversions directly on the website and not whether visitors visit the website and later decide to buy something in a physical store. The key measurement is the ratio of customers to visitors (i.e., the conversion rate). Due to limited resources and time, we did not conduct qualitative data collection and research, such as surveys, although it would have been desirable to do so.

1.8 Structure of the thesis

The thesis continues by giving some theoretical background concerning conversion, e-commerce, and HCI. This is followed by a description of the methods used to carry out the empirical study. Following this are details of the implementations and presentation of results, along with their analysis, and a discussion. Lastly, conclusions and recommendations for future work will be presented.

2 Background

This chapter provides background information about what a conversion is and working with conversion optimization in Sections 2.1 and 2.2. Sections 2.3, 2.4, and 2.7 mention elements of HCI design and the importance of: personas, scenarios, and general design principles. Sections 2.5 and 2.6 discuss exit-rates and how they are used in determining where to start the optimization process. Sections 2.8-2.11 reviews qualitative research concerning user psychology related to shopping. The sections compare online and physical shopping and focus on user preferences. Following this in Sections 2.12 and 2.13 is information concerning sampled data described together with the analytical tool used (Google Analytics). Lastly, related work is described and a short summary of the chapter are given in Sections 2.14 and 2.15.

2.1 Conversion

A conversion occurs when a visitor performs a measurable action that the e-merchant desires. This measurable action has an effect on the organization's overall results [6]. The term conversion is mainly used to describe the action occurring when a visitor makes a purchase and “converts” to a paying customer. A conversion can also, depending on the conversion goals, be the action occurring when a form is filled out, when a file is downloaded, or when a visitor clicks his or her way to a desired web page. Usually companies have several conversion goals that can be organized into either micro- or macro conversions. A micro conversion can occur, for instance, when a visitor puts a product in their shopping cart, whereas a macro conversion is when the visitor actually makes a purchase [7p. 13].

The proportion of conversions in relation to the total number of visitors is defined as the conversion rate; the higher the percentage, the better the conversion rate. E-commerce stores normally have a conversion rate on the order of less than 10% [8].

2.2 Working with Conversion Optimization

Knowing in which area to initiate the optimization process is one of the most important and toughest aspects of CRO. A mistake many designers make is to reflect solely upon the website's content and layout, rather than focusing on the most important fact—what one wants to achieve. Focusing upon *what* one wants to achieve should be the first step in CRO and then the focus shifts to *how* to achieve it [6].

Before examining and trying to understand a website's strengths and drawbacks, the e-merchant must first decide what conversion goals are to be achieved. These conversion goals vary between websites depending upon what they sell, their target group, and what the company as a whole wants to achieve. Goals

for e-commerce stores could be to increase the conversion rate for a specific product that is not selling well or for a specific target group that the business wants to reach. However, it is not until the goals are set that the process of implementing changes to increase the conversion rate can take place [6].

The next step is to gather knowledge of the business' users. How many people are completing the customer journey and what do they seek? Where along the journey are we losing or retaining customers? In order to answer our desired business questions we need to observe and measure the visitors' behavior and attempt to determine the visitors' wants and needs. With the knowledge gained from this data, a designer will be able to create a rational design hypothesis for every design change that they might want to make. This hypothesis should be based upon a reason as to why this change would increase the conversion rate, and preferably include an estimate of how much this change would increase the conversion rate. Having clear hypotheses facilitates structured testing and interpreting the results of this testing [4].

Many designers feel distress and annoyance over a low conversion rate and as a result, they eagerly change hundreds of factors at once in the hope that the changes have a positive effect. The drawback of making too many changes at once is that it becomes difficult to trace the factors that actually helped or hurt the conversion rate; therefore, it is important to set realistic goals for how much one wants the conversion rate to increase. The process of CRO takes this into consideration in the form of a belief that the cumulative effect of making small percentage increases over the long run will lead to success for the organization [4]. Iteratively making small changes is the key to identifying whether a given design change brings success or not. However, one of the most challenging parts of conversion optimization is to identify *where* the most impactful parts of a page are and deciding how to make the correct set of changes there.

It is also important to keep in mind during conversion optimization that even unsuccessful results should not be seen as failures. Instead one should seek to recognize what decisions have a positive or negative impact, and thus gain valuable information for further tests in the future. Even changes that have a negative effect on the conversion rate can give valuable information about your market and website [9]. However, there are reasons to be concerned when changes make no effect at all since this might indicate customers are not interacting with the website to a sufficient enough extent.

2.3 Getting to Know Your Customers: Developing Personas

A persona is a description of a fictitious person that is used to humanize and individualize a specific target group. These hypothetical individuals are used to *understand* your customers on a deeper level as they allow designers to build empathy towards them during the design process [3, pp. 106–111]. This will make it easier for designers to anticipate what questions such an individual will have and

where the site might be confusing by imagining the hypothetical user's likely facial expression [7, pp. 59–65].

In CRO, the data collected from online behavioral monitoring or from market research is translated into a set of personas. The benefit of personas can be seen when considering visitors of the different target groups, for which the overall goal is to appeal to individuals from every target group in order to capture their interest. Defining these individualized personas early in the design process helps create a design that is suitable for all of the target groups [7pp. 59–65].

2.4 Scenarios

Scenarios are stories in which the protagonists are the personas. Using scenarios, designers can place the personas into context and further bring them to life [3, pp. 144–145]. Scenarios are one of the essential factors that make personas worth having and they provide a fast and effective way to imagine design concepts in use. Using the same scenario with different personas provides a good technique for realizing what needs to be included in the final design stage [10, p. 359].

Consider our example of an e-commerce website. One persona is Stefan, a focused shopper who always knows exactly what he wants. Another persona is Karin, who likes to browse around and compare items. Imagining them in a scenario in which they are shopping for a product, the designer would in Stefan's scenario have him using search tools, whereas in Karin's case use browsing tools.

A common scenario designers' use is imagining the first-time use of a product or service by a persona. Questions such as "What will happen when the persona encounters the product or service for the first time?" and "How do they know what to do and how to use it?" arise, revealing how to tailor the final design to appeal to and work for each persona. Scenarios can take from a few minutes to an hour to write, whereas it takes significantly longer to storyboard, wireframe, and prototype.

2.5 Exit rate and using it to determine where to start to optimize

Exit rate is the percentage of visitors who exit the entire website from a specific page after visiting at least one other page in the website. The following formula is used to calculate the exit rate on a specific page (the formula is also used by Google Analytics) [7pp. 41–42]:

$$\text{Particular page exit rate} = \text{Number of page exits} / \text{Number of page views}$$

The exit rate is a useful metric early in a design process as it enables the website designer to determine where to start their optimization. By analyzing and evaluating exit rates of pages within the site, one can find faulty pages or other pages that need to be optimized by looking at how much they deviate from the average exit rate of the website. A mistake is to immediately start optimizing pages with higher exit rates than the site's average. Instead, the website should establish a

standard acceptable exit rate for various pages based on their function. According to K. Saleh and A. Shukairy, the following criteria have to be met for a page to be considered for optimization [7pp. 41–42]:

- The page has a higher exit rate than the acceptable exit rate based on the page's function.
- The number of unique page views is greater than X , where X is dependent upon both the size of the website and on the amount of revenue that can be generated by reducing the exit rate for that particular page.

2.6 Acceptable Exit Rates for a page

It should not come as a surprise that exit rates will vary between pages on a website. For example, it is natural that order confirmation pages or other completion pages have higher exit rates than other parts of the site since many users are expected to leave the site after making their purchase. Having an exit rate of 90% or more on such pages is typical. A general rule of thumb is that for pages from which visitors are expected to continue navigating, such as the product(s) or home page, the exit rate should be less than 10%-20% [7pp. 41–42]. Anything higher would be a good indication that there might be some hidden problems that need to be examined.

2.7 Designing for Interaction - Laws and Principles

The core of interaction design focuses on creating interfaces that are both engaging and well thought-out from a behavioral viewpoint. As understanding how users and technology communicate with each other is fundamental in the field, this knowledge can be used by the designer to anticipate how users might interact with the system. The process of fixing problems early on and inventing new ways of doing things become much easier.

As interaction design is a fairly new field there are no rules or “laws” set in stone. Although interaction designers are still figuring out the basic principles of the work they do, there exist a handful of laws these designers use [3, pp. 129-138]. However, these laws and principles should *guide* the designer and not dictate the design.

Table 2-1 shows some questions to consider when designing for interaction and the principles related to them.

Table 2-1: Principles [3, pp.129-138]

<p>Define how users can interact with the website</p> <p>Entry: 1</p>	<p>What can a user do with their mouse, finger, or stylus to directly interact with the website?</p> <p><i>1. Direct Manipulation</i> Mimicking an action we might perform on a similar object in the physical world, for instance, to drag and drop, resize the window, and pushing buttons. Because such actions closely map to our physical experience, these types of direct manipulations supposedly make an interface easier to learn and use, especially for 3-D objects in a digital space.</p>
<p>Give clues about behavior before actions are taken</p> <p>Entry: 2</p>	<p>Does the appearance (color, shape, size, etc.) give users a clue about how it functions?</p> <p><i>2.1. Affordances</i> Consider using properties to provide some indication of how to interact with an object or feature. Appearance is important and we want users to discover and use the functionality of a product in a correct manner. For example, “you know you can push a button because you have pushed one before” [3, p. 131]. For instance, let the increment/decrement button have a “+” and “-” symbol.</p> <p>What information is provided to let a user know what will happen <i>before</i> they perform an action?</p> <p><i>2.2. Feedforward</i> Letting users know what will happen before performing an action gives confidence. For instance, you can provide instructions before a final submission or use meaningful labels such as “Pushing this button will do that”.</p>

<p>Anticipate and Mitigate Errors</p> <p>Entry: 3</p>	<p>Are there constraints to help prevent errors?</p> <p><i>3.1 The Poka-Yoke Principle</i> Putting constraints on products to prevent errors, forces users to adjust their behavior to correctly execute an action. Implications of this in interaction design occurs when designers disable functionality (or the navigation, menu items, or the icon) when conditions for its use have not yet been met. This ensures that proper conditions exist <i>before</i> a process begins, preventing problems from occurring in the first place. For example, constraining users to decrease the quantity of items below 1.</p> <p>Do the error messages provide a way for the user to correct the problem or at least explain why the error occurred?</p> <p><i>3.2 Errors</i> Provide users with a way to fix the error, or at least provide information about why the error occurred.</p>
<p>Consider System Feedback and Response Time</p> <p>Entry: 4</p>	<p>What feedback is given once a user performs an action?</p> <p><i>4. Feedback</i> Feedback gives an indication that something has happened (i.e. some notification). Feedback should occur early and often, as it is important for the user to get an acknowledgement from the system.</p> <p>Do we know that the product has “heard” what we have told it?</p> <p><i>4. Feedback (continued)</i> Providing a mechanism that lets users know that the system has heard their request and is working on it is a good design principle. Psychologically speaking, this makes the waiting period seem shorter even though it is not. For example, instead of using <i>spinning wheels</i>, tell the user what is happening when installing software.</p>

<p>Strategically think about each element</p> <p>Entry: 5</p>	<p>Are you following standard design conventions?</p> <p><i>5.1. Standards</i> “Obey standards unless there is a truly superior alternative” [11] is a well-known quote in interaction design. Freely propose new methods, but do so with care as these new methods subverts the user’s expectations of how a product should work. Throughout the years designers have trained users to expect certain elements to be located in certain places (for instance, by placing the company’s logo at the top left of the website). Making users learn something different can cause distress.</p> <p>Are the interface elements a reasonable size to interact with?</p> <p><i>5.2. Fitts’s Laws - Create Larger Targets & Minimize Cursor Movements & Avoid Muscular Tension</i> Fitts’s law states that there are two things that determine the time it takes to move from a starting position to a final target: the distance to the target and the size of the target. The bigger the target, the faster it can be pointed to. The closer the target, the faster it can be pointed to. Make buttons reasonably big and close to the relevant elements; this is especially important when using mobile devices and touchscreens, to minimize muscular tension. For instance, a horizontally designed check-out page allows users to avoid muscular tension in terms of scrolling.</p> <p>Are interactive elements, such as menus, strategically placed at edges and corners?</p> <p><i>5.3. Fitts’s Law - Exploit The Prime Pixels</i> No matter how far one tries to move the cursor, it will always stop on the edge and land on the menu. Positioning menu bars and buttons at these locations is an excellent choice as edges and corners have infinite height or width, and require no mouse precision to find.</p>
<p>Simplify for Learnability</p> <p>Entry: 6</p>	<p>Is information chunked into seven (plus or minus two) items at a time?</p> <p><i>6.1 The Magic Number Seven</i> The human mind is optimally able to remember information in their short-term memory in chunks of 7 before making errors. Designers often mistake the implications of this by never having more than seven items on a screen at once, but it is important to know that this number concerns information that one is forced to remember in short-term memory. The lesson designers should take from this is to not design a product that causes “cognitive overload” by ignoring the rule.</p> <p>Is user’s end simplified as much as possible?</p> <p><i>6.2 Tesler’s Law of the Conservation of Complexity</i> This law states that there is a point beyond which you cannot reduce the complexity any further; hence you can only move the inherent complexity from one place to another (perhaps to the software). Try to remove as much complexity as possible from the</p>

	<p>user and instead design in a way that the system does as much work as possible. As an example, implement increase/decrease button instead of users manually type in the quantity.</p> <p>Are familiar formats used?</p> <p><i>6.3. Hick's Law</i></p> <p>Hick's Law states that users' time to make decisions is affected by the number of possible choices they have. It also states that the two factors affecting the decision time are: how familiar they are with the choices and the format of them. This occurs because users subdivide the choices into categories and eliminate nearly half of the remaining choices with each decision step. As an example, the more options a user has to pick from—be it navigation or products to look at, the more energy it takes to make a decision. In the end, the energy required becomes so large that the benefit of making a decision does not seem worthwhile.</p>
--	---

2.8 Important features of an e-commerce store

In a study conducted by HUI Research in collaboration with PostNord and Svensk Digital Handel, Swedish consumers in the age range 18-79 were surveyed on their behaviors, opinions, and habits while e-shopping. One question was “How important are the following characteristics regarding the shop's layout and information when deciding which Web store to shop from?” [1, p. 40]. The responses are shown in Table 2-2.

Table 2-2: Important features regarding website's layout, according to customers [1, p. 40].

Important Layout Feature	Share that considered it important
1. Clear Product Information	92%
2. Total Price	90%
3. Easy to navigate	83%
4. Contact Customer service	81%
5. Secure e-commerce certificate	62%
6. Customer Reviews	42%
7. Responsive design (mobile)	27%

2.9 Strengths and Weaknesses of Physical and E-commerce Stores

Today's customers face the choice of whether to purchase in a physical store or an online store. However, the driving force of purchasing seems to differ between them. The reasons why customers chose to buy a product in a physical store, as opposed to an online store, are shown in Table 2-3. These answers are based on people who have recently made a purchase in a physical store (80%) whilst the reasons for purchasing online are listed in Table 2-4. These answers are based on people who have recently made a purchase in an e-commerce store.

Table 2-3: “Why I purchased in a physical store instead of an online store”
[1, pp. 42–43]

Comfort suited me better	30%
Too long delivery time	29%
Want to test and feel the product	27%

Table 2-4: “Why I purchased in an online store instead of a physical store”
[1, pp. 42–43]

Cheaper	31%
Comfort - shopping when it suits me	31%
The supply is not local	21%

Participants in the study felt that their chosen medium of purchase was more comfortable than the alternative. Table 2-4 shows that the strength of e-commerce shopping is product price and lack of local availability of the product. However, Table 2-3 shows that customers prefer to shop in a physical store when it is necessary to test and feel the product along with avoiding waiting for delivery.

2.10 Consumers and Companies' view of them

The factors that play important roles for a visitor in the process of shopping in an e-commerce store are described in this section. Before discussing the points one by one (except for those that cannot be changed by design), we can see from Figure 2-1 that corporations seem to misjudge how important visitors think the search function and the price are [1p. 43].

Based on participants - both consumers and companies- that had e-shopped, 93%; answered the question: "How important are the following factors when shopping from an online store/web page?" (Share that answered "very important" or "important")

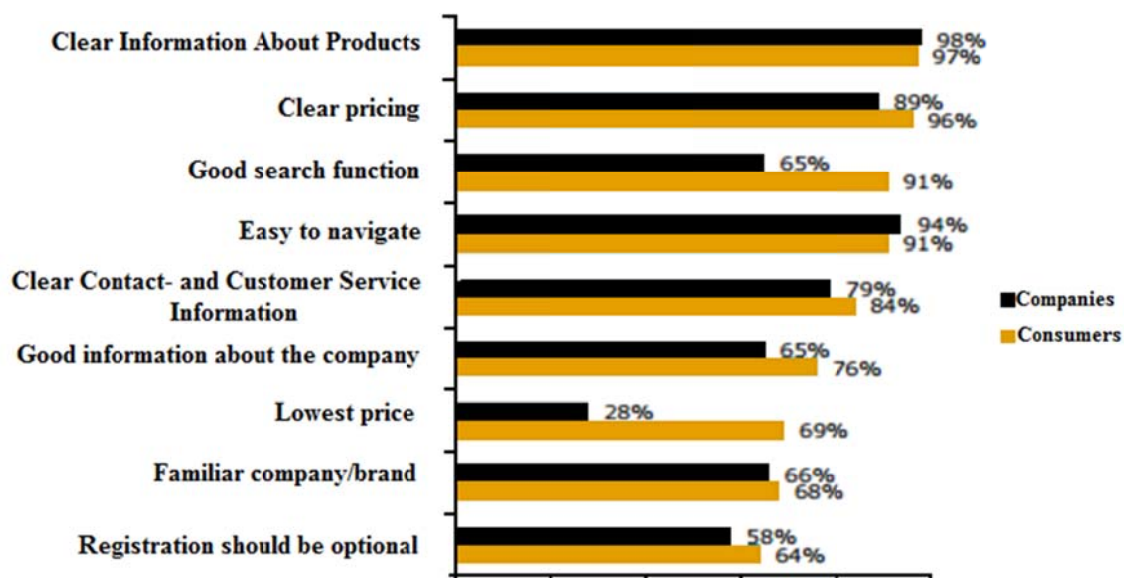


Figure 2-1: Survey answer distribution among consumers and companies to the question: How Important are the following factors when shopping from an online store? [1p. 43]

2.10.1 Clear Information about Products

It is important that the products that e-commerce tries to sell are presented clearly. One drawback of online-stores compared to physical ones is that the customer lacks the ability to feel and test the product before purchasing. Therefore, it is essential to offer plenty of product information including high-resolution images of the product, preferably with views from several different angles.

2.10.2 Total price

One reason why visitors abandon their shopping cart is because the total cost was not listed until the end of the check-out process. While the price cannot be affected by design measures, keeping the total price clearly visible at all times rather than leaving it to the very end will avoid user frustration.

2.10.3 Good search function

An important aspect of a website is the search function. The website's built-in search engine must not only work without the user experiencing any difficulties, but the search engine must also generate appropriate direct references to the desired information. A search that does not generate any results should not display an empty page but instead, should display tips about popular/related/available offers. It is also important to know what searches occur most frequently so that in these cases the company can promote those products and offers that are most important to the business [12].

2.10.4 Easy to navigate

Having consistent and predictable navigation is important in order to avoid invoking (in the visitors) a fear of getting lost. Visitors should know where they currently are and understand the facilities to move around. Not only does poor navigation make visitors confused and frustrated as they try to figure out how the website works, but it also can have a bad impact on conversions since customers who cannot find what they are looking for, simply cannot buy it [13pp. 184–188].

2.10.5 Clear Contact- and Customer Service Information

An important factor in being perceived as a serious e-commerce store is to have clear contact and customer service information. In order to optimize the level of trust for the user, visitors to a site should be able to easily find information about the company and its employees, preferably with images and a short description of the individuals. Therefore, it is important to show visitors how to contact customer service and make it simple for visitors to find this information [13pp. 340–346]. Moreover, in an e-commerce survey [1p. 46] when companies were asked what factor they think is most important to increase the conversion rate, 73% said “that the customer needs to feel the website is secure”.

2.10.6 Registration should be optional

There are several reasons why visitors abandon their purchase, one of this is due to forced registration [14]. One reason for users avoiding registration is that users already have a large number of usernames and passwords to remember and they do not want to create an entirely new account just to purchase an item or two. Another reason is the expectation users have of receiving junk mail containing marketing material. A third reason is that signing up for an account takes time; hence it goes against Tesler's law of conserving complexity. Other reasons include the confusion of why an account is needed to buy a product and because of the realization that the website is going to store their personal information indefinitely—giving users an uncomfortable feeling.

2.11 Differences in Male and Female Consumer Behavior (in e-commerce)

There are considerable gender differences in consumption patterns. There are 3 times more women than men who buy clothes and shoes, whereas there are 3 times more men than women who buy electronics [15, p. 17]. Similarly, there are 3 times more women than men who shop for home decor and furniture [1, p.11]. However, other purchase frequencies are more or less the same between the genders.

The behavioral patterns between males and females are quite contrary to one another when comparing a physical store with an online store. In physical stores, women demand environments in which they can move freely at their own pace and take the time they need to find the product they are searching for. Impulse shopping along the way is usual. However, for men time-efficiency is their focus, hence they prefer environments where they can find the product they are searching for in the shortest possible time and then leave as fast as possible [2].

The genders switch behavior when it comes to e-commerce, as it is males who spend time searching for different products via different pages, while females are time-efficient and tend to terminate their web search as quickly as possible once they have found one source to buy from [2].

2.12 Sample Data

A problem faced when trying to use data based on a lot of visitors, or a large population, is that making a census or a complete listing of all the values in that population is either impractical or impossible. Therefore, in statistics one usually selects a sample of a large population, i.e., a subset of manageable size, when making inferences or extrapolations. Sampling data in such way is widely used in statistical analysis when the analysis of a subset of the data gives similar results to an analysis of the complete data set. Google Analytics, the web analytic tool in use, automatically samples the data when more than 500 000 sessions are collected, allowing rapid return of results due to reduced processing time [16].

2.13 Google Analytics

Google Analytics is freemium web analytics tool used to track and report website traffic. It is used by 66.2% of the 10 000 most popular websites and is continuously being expanded with additional functionality [17]. With respect to e-commerce, Google Analytics can track and report a website's transactions, revenue, and many other commerce-related metrics [18]. Since our aim is to increase conversions on the website, we measure progress by observing the following Key Performing Metrics: conversion rate, the number of transactions, and revenue.

2.13.1 Visitor Segmentation

Segmenting (filtering) the data is a vital tool to use in Google Analytics when trying to find insights concerning visitors or sessions. For example, one can filter the data by choosing users who are male, female, or who have made a purchase in order to discover new insights about members of that segment. After discovering insights about a particular segment, one can then find the best way to improve their value. Segmentation helps the analyst understand the customers and segmentation is commonly used as a way to identify and prioritize those target groups that the company aims to improve conversion for, hence the segment chosen normally includes high-value customers [19].

2.13.2 Sales Performance

Google Analytics' sales performance tool gives the user an overview of how sales are going for all products. For every product the tool lists: how much revenue, how many unique purchases, and the quantity of purchases that have been made. This can give the user valuable insights as to which product pages actually make an impact on the overall website. Then these product pages can be further investigated, hopefully leading to where one should initiate the optimization process.

2.13.3 Sessions

It is important to understand the concept of a session in Google Analytics since many reports and metrics depend on how Analytics calculates what a session is. User sessions ought not to be confused with page views, since a single session can contain multiple page views, events, and e-commerce transactions. A session can be considered as a container for those actions a user makes on the website. By default, sessions last until the user has been inactive for 30 minutes. Additionally, sessions can be as short as a few seconds if the user chooses to exit the website, or as long as several hours assuming continuous interaction [20].

2.14 Related work

This section describes the related work others have produced concerning CRO and e-commerce.

2.14.1 Background on conversion optimization

Saleh and Shukairy [7] emphasize the importance of conversion optimization in their book. They describe all aspects of conversion optimization, ranging from how to attract users to a page to how to make loyal return customers. They also touch upon ways to start the optimization process if one cannot afford to conduct qualitative research, namely, through inspecting high exit-high value pages.

2.14.2 Importance of knowledge about target groups

Thörn [43], whose focus is mainly on increasing conversion rates through performing usability testing with users from the website, claims that the key to identifying the strengths and weaknesses of an e-commerce store lies in having as much knowledge as possible of the main target groups. He further claims that having direct contact with the website's main users is necessary in order to create an appropriate design.

2.14.3 Design areas affecting conversion rate

Lundvall [44] wrote about conversion rates and usability, claims that there are 3 main design-related areas that affect the conversion rate: layout, navigation, and trust in the company. Additionally, he concluded that one has to achieve a synergy between external factors, design-related factors, generalization issues, and the importance of testing iteratively in order to create a long-term conversion workflow.

2.15 Summary

To summarize, this chapter emphasizes on the importance of; knowing your users before making a design, figuring out where the site needs an improvement or where a design change can make a significant improvement, users wants' and needs' in that specific market, and the most important aspect when working with conversion optimization – to only make one change at a time and to determine its effect. The reason behind collecting all these types of data, as well as behavioral science data of different target groups interacting with such website, is to provide the designer with all necessary tools that are required to produce a more optimal design.

3 Methodology

The purpose of this chapter is to provide an overview of the research method used in this thesis. Section 3.1 describes the research process. Section 3.2 details the research paradigm. Section 3.3 focuses on the data collection techniques used for this research. Section 3.4 describes the experimental design, the choice of method, its benefits and limitations, as well as techniques used to evaluate the reliability and validity of the data collected. Section 3.5 describes the method used for the data analysis. Finally, Section 3.6 describes the framework selected to evaluate our method.

3.1 Research Process

To accomplish the goals of this thesis project, a combination of processes are used in conjunction with one another. It is important to realize that the data collection process aimed to identify two main issues: (1) who our website's users are and (2) where to start the optimization of the website. The A/B testing process aimed to identify how we measure success/failure, how to generate hypotheses, and to create and run the experiment. A third process concerning evaluating of the pages to optimize had design flaws when considering important e-commerce and design factors valued by users and general HCI design laws & principles. We inserted additional procedures into a standard A/B testing framework [21]. Details of the complete process are described below:

1. **Collect Data (using Google Analytics):** The data collected from the qualitative research and/or quantitative research provides insights into where the optimization process should begin. The recommendation is to test highly trafficked pages, since that will allow more rapid data collection. It is also a custom to inspect pages with either low conversion rates or high drop-off rates.
 1. Getting to know the websites main users (leading to values to be inserted into the Personas):
 - a. Demographics (percentage gender breakdown of both visitors and converts),
 - b. Interests & Hobbies,
 - c. Device used,
 - d. Time of day when visiting the page, and
 - e. Location

leads to development of personas & scenarios.
 2. Where to start the Optimization? Which page(s)?
Inspect high valued and highly trafficked pages.
 3. Pages that have made a turnover of over X SEK.
Check the exit rate of each page (if >20%?)

4. Inspect the check-out funnel page(s). This area is critical as the users are one click away from placing an order.

Check the exit rate of that page (if <75%?)

2. Does the desired page(s) disobey design laws, principles, and/or e-commerce values important to users?

Inspect page elements using the designing for interaction principles listed in Table 2-1, the important features of an e-commerce store in Section 2.8, and those factors that play an important role for visitors when shopping in an e-commerce store in Section 2.10.

3. **Identify Goals:** A test should have clear conversion goals, as these are the metrics used to determine whether or not the variation is more successful than the original version. The goals can be straightforward, such as having a visitor: clicking a button, link to product purchases, or sign up for an e-mail newsletter.
4. **Generate Hypothesis:** After the goals have been identified one should begin to construct A/B testing ideas and hypotheses in the form of “If [variable], then [result] due to [rationale]”. The variable is the website element in question that can be added, modified, or deleted to produce a desired outcome. The result is the predicted outcome, i.e. more purchases or more email sign-ups, etc. The rationale demonstrates that the reasoning behind your hypothesis is supported by research concerning what you know about your visitors and their behavior. Well-thought out hypotheses makes it easy to interpret the results of a test even if the hypothesis proves to be incorrect. The important fact is that you motivate why a change will be better than the current version.
5. **Create and run the experiment:** Most A/B testing tools have a visual editor that assists in making changes and implementing them. Once these changes have been made you run the experiment and wait for visitors to interact with the website. Visitors’ interactions with each variant of a page are measured, counted, and the results are compared to determine how the original and variant perform.
6. **Analyze Results:** Once the experiment is completed, it is time to analyze which version was better and/or worse and what lessons can be drawn from this. The results are evaluated in relation to the hypothesis in order to draw conclusions about why the assumption was correct or incorrect. Unsuccessful results in which the desired goals were not met can lay the foundation for interesting conclusions.

3.2 Research Paradigm

This work follows a so-called post-positivistic view that context is necessary for our data and research to be considered realistic or true [22]. This approach is suitable for our experimental methodology: a statistical and quantitative study of the subject. However, it is also necessary to have an approach based on constructivism. This is because we must account for the possibility of multiple views, opinions,

preferences, interpretations, etc. A web design is very subjective because of its opinion based nature. This is the ontology commonly used for qualitative research that we will use to guide us in approaching changes to the website's design.

3.3 Data Collection

This sub-section describes the web analytics- and A/B testing tools we used, namely, Google Analytics and Optimizely. It is described why these tools were used, along with social and ethical concerns these can bring. We also discuss whether sampling took place in our data collection and A/B tests. We also touch upon demographics and target group(s).

3.3.1 Google Analytics

Google Analytics was used as the data collection tool to complete the first procedure in the research process enumerated in Section 3.1. The tool was described in Section 2.13 and was the obvious tool of choice due to Nordic Design Collective already having implemented it and having used it to evaluate the website for the last two years. This allowed us to immediately analyze data, rather than needing to wait to collect data. As described in Section 2.13, Google Analytics is the most popular web analytics tool and additionally, unlike other analytic tools, it uses google-account information to identify each visitor's age and gender when they access the website. However, information that identifies an individual person is not permitted in Google Analytics. If the website using Google Analytics does collect personally identifiable information, then it violates Google's terms of service and Google is allowed to delete all data that has been collected.

According to the Electronic Communications Act [14], it is required that a website with cookies provide all visitor's with information that cookies are enabled and what they are used for. Nordic Design Collective does this in order to comply with this law. Their use of cookies is for tracking sessions and for Google Analytics data collection. They do not disclose any personal information to any third parties. Users are mostly unaware of their browsing being monitored due to not reading the cookie policies document. Moreover, their behavior is *not* publicly shared.

3.3.2 Optimizely

There are different web solutions available to perform A/B testing, such as Optimizely, Visual Web Optimizer, and Google Analytics - Content Experiments. We selected Optimizely, not only because it has a simple to use a visual-based editor (unlike Google's "Content Experiments") but also because the company already had a software license for Optimizely. Optimizely is the world's leading digital experimentation platform and has delivered over 700 billion experiences tailored to the needs of the customers of marketers, developers, and product managers worldwide [23, 24]. Unlike many other A/B testing tools that use

traditional fixed hypothesis testing in determining validity of results, Optimizely switched its statistical engine to a more suitable method for today's digital experimentation. As described in Section 3.4.8, Optimizely supports sequential testing and false discovery rate (FDR) control, thus avoiding many pitfalls experimenters face with traditional hypothesis testing and offering better error-rate control.

3.3.3 Sampling

Sampling was described in Section 2.12. Google Analytics did not perform sampling in our case, since there were less than 500 000 users who accessed Nordic Design Collective's website during the time frames used when evaluating data. The lack of sampling also indicates that the datasets were small.

3.3.4 Sample Size

For Optimizely there was no specific predetermined sample size we were trying to reach. This was acceptable because Optimizely determines statistical significance using FDR, rather than the Type-1 error rate used in traditional statistical testing (as discussed in Section 3.4.8).

3.3.5 Demographics and Target Groups

We used Google Analytics to determine the website's demographics. This was beneficial as there is a lot of data collected by Google's other services that would otherwise be unavailable to us without requiring a lot of qualitative research. This motivated our choice of Google Analytics. By default, Google Analytics shows the entire website's population. However, by filtering the view using demographics, interests/hobbies, the device used, time of day, location, etc., one can tell what kind of people visit the website. This data reflects our population. However, Nordic Design Collective's view of their target group(s) may not actually reflect the demographic of the website visitors that they attract. This is why it is necessary to collect evidence through the use of data collection tools.

3.4 Experimental design/Planned Measurements

Section 3.4.1 compares qualitative and quantitative research. Sections 3.4.2- 3.4.4 describes and compares A/B testing with other testing methods, as well as mentions the benefits and limitations of it. Section 3.4.5 mentions the test environment, components needed to reproduce the tests. Sections 3.4.6-3.4.8 describes the reliability, statistical significance, and the validity.

3.4.1 Qualitative versus Quantitative Research

Qualitative studies are typically used to gain a better understanding of the target population's views and reasons for these views, seek answers to questions, and provide evidence of why visitors behave as they do. These studies can take the form of questionnaires, usability testing, focus groups, etc. Conversely, in quantitative research one investigates observable phenomena via statistical, mathematical, or computational techniques. A quantitative approach allows the researcher to develop and employ hypotheses pertaining to a phenomenon, as this hypothesis tends to answer the question of *how* a population behaves as opposed to *why* they behave in a specific manner. It is best practice to use both qualitative and quantitative approaches in a study in order to find where to optimize a website and to gain insights about different target groups, while using qualitative research with specific users to answer questions, such as why they behave as they do, what aspects of the website can be improved, etc.

Since time and resources were limited, we were unable to conduct qualitative research on our own. Instead, we compensated for this by using e-commerce related qualitative studies done by HUI research, PostNord, and DHL, as described in Sections 2.8-2.10. These qualitative studies were purposefully chosen because these surveys were conducted on people living in Sweden, the region from where over 78% of Nordic Design Collective's traffic comes. Our design hypotheses were formed using the results of these earlier surveys. Additionally, these studies had a bigger sample size than would have been possible if we had conducted a study based only upon the visitors to Nordic Design Collective's website. This was necessary because the number of visitors to their website is quite small; hence the actual sample size is insufficient to reflect their entire desired population.

3.4.2 A/B Testing vs Multivariate Testing vs Multi-Armed Bandit Testing

There are several techniques available to perform tests of a web page, such as A/B testing, Multivariate testing, and Multi-armed bandit testing. These techniques will be compared below, but the main reasons why A/B testing was selected was our prior knowledge of the low number of visitors to Nordic Design Collective's website and the importance of achieving statistically significant results.

Multivariate testing is an approach used for testing a hypothesis in which multiple variables are modified. In this type of testing the experimenter wants to test several variations to elements, with the goal of determining which variations perform best out of all possible combinations [25]. The problem with this approach concerns the minimum amount of traffic required to reach meaningful results. In A/B testing, traffic is split evenly with 50% to the original version of the website and 50% to the variation. In multivariate testing, traffic will be split into smaller segments, thus each variant will receive a small portion of the traffic. This can greatly increase the duration of an experiment, something that was unwanted. According to Leonid Pekelis, another difficulty of multivariate testing is the risk of

more false positives, since each testing variant has a 5% rate of finding a false positive [25]. He further claims that there are ways to account for this, but the cost is the need for even more visitors to achieve conclusive results.

Another technique is the Multi-armed bandit test, in which two or more variations run simultaneously, initially with an equal amount of traffic (during 10% of the experiment's duration). The remaining 90% of the time, traffic is automatically allocated to the currently best-performing version [26]. This 10/90 ratio can be tweaked, but it is important to note that it early on sends traffic to the currently winning variation, allowing the average conversion rate to be higher than for an A/B test in which an equal fraction of traffic is sent even to a worse performing variant. It is important to realize that the fundamental concept of statistical significance is missing in this technique, as we do *not* decide which variant to allocate more traffic to based on a significant change in the number of visitors. This means that there is a risk of losing some sales and conversions with A/B testing, but this is the price you pay for *finding out if* the variation really performs badly or not. With A/B testing there is a certainty as to whether one variant is really beneficial or not. With Multi-armed bandit testing, a little traffic early on when deciding to which alternative to allocate traffic leads to a lot of uncertainty as to whether the variant really performing worse than the current version or not. One can adjust the above-mentioned ratios in bandit tests, but the need for a considerable number of visitors remains [26].

3.4.3 A/B Testing

A/B testing is a method in which two versions of the same web page compete against each other by exposing 50% of visitors randomly to version A (the control) and 50% to version B (variation). Statistical analysis in form of hypothesis testing (described in more detail later) is used to determine which variant performs better with respect to a certain conversion goal [7p. 195].

Although A/B-testing can have many goals, it is usually done to obtain concrete results of how design changes affect conversion on a website. Well thought through tests can give important insights about which design decisions improve a website and which ones affect it in a negative way. A/B testing takes the guesswork out of optimization and enables data-informed decisions, thus shifting conversations from “we think” to “we know”. Today large companies perform continuous A/B testing to improve their website's design and increase their conversion rate [27p. 217].

Performing random tests and hoping for a good outcome is not recommended. Generally, this is considered a waste of both time and money. In contrast, one should build design proposal hypotheses through usability testing, expert evaluations, web analytics data, and/or from previous hypotheses. The best-case scenario would be to use all of the mentioned parameters when making design proposals; however, since this method takes a lot of resources. It is nonetheless

considered a bad practice to use only one of these parameters. The least expensive approach is to use a free web analytics tool, such as Google Analytics, to retrieve valuable data - such as audience demographics and audience behavior. For example, the metrics of common exit pages and the exit rates of those pages can be helpful in determining where the problem may lie.

It is important to realize that tests resulting in improvements do not mean that the work is complete and that the results will remain positive. Confounding variables and external factors cause the data to be non-stationary. Stationary data occurs in a time series in which the statistical properties (mean, variance, autocorrelation, etc.) are constant over time. This lack of stationarity needs to be taken into consideration, especially for e-commerce stores since we cannot make the same assumptions as we could with stationary data. Some reasons why the results might fluctuate include [28]:

- Season,
- Day of the week,
- Holidays,
- Press (positive or negative),
- Pay per click,
- Passing of information by oral communication (word of mouth),
- Search engine optimization, and
- Newly formed trends.

In addition, there are many causes for fluctuation in results. Keep in mind that these fluctuations do not mean that the data is unreliable. However, since both version A and B are exposed simultaneously to 50% of visitors, it is possible to identify trends. This suggests that one should not compare the results with data from previous periods/months when an A/B test was not running. Alex Birkett recommends running a follow-up test during the oncoming period [28].

3.4.4 Benefits and limitations of A/B-tests

There are several benefits and limitations of A/B testing as a research method. It is important to remember that A/B testing answers the question of *how* users behave and not *why* they behave as they do. The benefits and limitations of A/B testing are summarized in Table 3-1.

Table 3-1: Benefits and limitations of A/B-tests [29]

Benefits	Limitations
<p>A/B-tests measure users' actual behavior and can be seen in real-time as the test is running, making it easy to determine which version performs better than the other. One can then confidently conclude that the better performing version is the one that should be shown to all users in the future.</p>	<p>A/B-testing can only be used for projects that have one clear some Key Performance Indicator (KPI) that is measurable by a computer. For instance, not all websites have a measurable user action (such as sales for an e-commerce site or subscribing to email newsletters).</p> <p>As some KPIs only measure a single desired action from a visitor, one cannot ensure that the action in question is the cause of a higher conversion rate. The visitor's decision to convert may depend on several different factors that cannot be measured during an A/B test.</p>
<p>A/B-tests replaces the “we think” guesswork with the “we know” how design changes affect users by confirming design proposal decisions with on-site user engagement, the number of visitors, and conversion data as measured with high statistical significance (assuming that the tests are exposed to a sufficiently large number of visitors).</p>	<p>A/B testing is complicated and time consuming when it comes to creating and fully implement different test versions in the current interface. Many e-commerce companies do not have full control of their website's source code, making A/B testing suitable only for a very small number of ideas.</p>
<p>A/B-testing is a cheap method and it is free of charge to collect and analyze the data using various web solutions found online.</p>	

3.4.5 Test environment/test bed/model

In order to reproduce our test environment, the experimenter needs to have a website to experiment on. This website needs to be linked to both a data collection tool (such as Google Analytics) and an A/B testing tool (such as Optimizely). The website in question must (as described by the A/B-testing limitation 1 in Section 3.4.4) have one clear KPI goal that is measurable by a computer, such as a “placing order” button or subscribing to a newsletter button. Furthermore, it would be beneficial if the website has a high visitor flow, sufficient to make a decision based on a statistically significant result. Having a high flow of visitors will reduce the time to complete tests in contrast to a low visitor flow scenario. The experimenter

should also research modern design laws, principles, and what the current users' wants and needs are for the particular category of website when generating their design hypothesis.

3.4.6 Reliability

It is necessary to know somewhat how Google Analytics and Optimizely are implemented in order to better understand the causes of inaccuracy in online data collection. Both Google Analytics and Optimizely are implemented using snippets of JavaScript tracking code which the webmaster adds to every page on the website that is to be tracked or experimented on [30][31]. This code is placed in the page's header (i.e., within the <head></head> tags). This code will tag a visitor with a cookie which allows visitor behavior data to be collected which is then returned to Google's and Optimizely's servers. The cookie is stored as a file in each visitor's device and it is used for websites to identify visitors and their purchasing habits.

This industry-standard method of JavaScript embedded "page tagging" yields reliable trends and a high degree of precision [32], but it has its limitations. The data collection can show inaccurate results due to:

Users deleting or blocking cookies	Web analytics depends on cookies to identify unique visitors in their statistics by using a persistent cookie that holds a unique visitor ID. After deleting these cookies, the user will appear as a new first-time visitor at their next interaction point, reducing the accuracy of conversions, click-stream analysis, and other metrics that depend upon the activities of a unique visitor over time.
Users having ad filtering programs and extensions	Ad-blocking and script blocking extensions can block tracking codes and prevent some traffic and users from being tracked, leading to holes in the collected data.
Users browsing through anonymity networks	Privacy networks such as Tor will mask the user's true location and present geographical data that is inaccurate.
Users having JavaScript disabled in their browsers	Data collection tools using "page tagging" such as Google Analytics and Optimizely cannot collect data unless the user's browser has JavaScript enabled, as the tracking codes are implemented using JavaScript.

Multiple users on the same device	Since the cookie is set only once in the device, the analytics tool will not spot a difference in whether or not someone else is using the same device when interacting with the website. This should be counted as two separate unique visits, but will be counted as one unique visit due to the cookie being tied to one device.
The same user using multiple devices	Researching products on a mobile device, but later on buying it through another computer will attribute the purchase to a brand new visit. In a perfect world, the cookie stored in the mobile device would allow the behavior to be tracked even when one switches device.
Sampling data	Although in our case Google Analytics and Optimizely did not use sampling when reporting results, as mentioned in Section 2.12 - Google Analytics samples data after reaching a threshold of 500 000 visits or views.

3.4.7 Statistical Significance

It is important in any quantitative study to perform statistical analysis to conclude whether the results are due to random chance or not. The result of an experiment is said to be statistically significant if it is likely not caused by chance for a given statistical significance level. Statistical significance is important in A/B testing since it gives the experimenter and the company confidence that the changes they made to the website actually have a positive or negative impact on the conversion rate [7, pp. 41–42].

The default methodology used to evaluate whether results are significant or not is the null hypothesis, in which experimenters assume that their variation will perform the same as the original. The goal of the hypothesis test is to try to disprove the null hypothesis and to answer the conditional probability question “Given that there is no change, what is the probability of obtaining the observed (variation conversions) data?”. As the standard level of declaring results significant in statistics is 95%, the experimenter checks whether the observed data has less than a 5% probability of being obtained by chance. If so, then we reject the null hypothesis and conclude with 95% confidence that the impact of the change is not random.

The two key variables that affect the significance level in A/B tests are the number of visitors and the fraction of them that convert. A/B testing a page that initially has a low baseline conversion rate and shows low improvement (i.e., a small effect) will require more visitors until the improvement is considered significant. Likewise, the higher the baseline conversion rate and the larger the improvement, the fewer visitors needed. A test should normally continue to run until the results have reached a level of 95% significance; however, in some cases it

is justifiable to accept an 80% level of significance if the company cannot afford to wait any longer [7, p. 198]. However, one should in such a case be cautious in implementing the variation since the results are more likely to be caused by chance as compared to a 95% level. When these levels of significance are not shown, the results should be considered inconclusive.

3.4.8 Validity using Sequential Testing combined with FDR control

As of 2015, Optimizely has shifted its statistical calculation engine process from traditional, fixed horizon hypothesis testing to a process combining sequential testing and FDR control [33]. This new statistical framework for A/B testing seems more suitable for today's digital experimentation and avoids many pitfalls experimenters were exposed to with traditional A/B testing statistics, such as:

- Setting a minimal detectable effect and sample size in advance is inefficient and non-intuitive
- Continuous monitoring (peeking at your results before reaching a predetermined sample size) can introduce errors into the results and cause you to take action based on false winners (Type 1 Error), and
- Testing a larger number of goals and variation at once greatly increases errors due to false discovery (the “multiple testing problem”).

The limitations of fixed horizon testing constrain the experimenter, as it assumes that evaluation of the experimental data will only occur at one point in time, at a set sample size. Experimenters rarely have a fixed sample size or a sense of the minimal detectable effect the variation will make in advance. Therefore, sequential testing is more effective as it is designed to evaluate experiment data as it is collected. The tests can be stopped at any time, while still giving valid results.

Optimizely's implementation of sequential testing calculates an average likelihood ratio – the relative likelihood that the variation differs from the baseline every time a new visitor triggers an event on the page. The p-value (which helps you determining the significance of your results in traditional testing) now represents the likelihood that the test will ever reach the desired significance threshold that you chose [33]. One can think of this as a traditional p-value for a world in which the sample size is dynamic. The process is called “a test of power one” and is better suited than traditional t-tests for the objective A/B testers.

An example of how much error rate will be added every time the experimenter “stops and peeks” on an ongoing test between a traditional fixed horizon tests and sequential tests is shown in Figure 3-1. The error rate in question is Type 1 error, the incorrect rejection of a true null hypothesis (a “false positive”).

Intervals of monitoring visitor data								
	500		1 000		5 000		10 000	
Traditional Error Rates	5%	+	5%	+	5%	+	5%	>5%
Sequential Testing Error Rates	1%	+	0,50%	+	1,50%	+	1,50%	<5%

Figure 3-1: Difference of increase in Type 1 Error rates between traditional and sequential testing when continuously monitoring (4 times) [34].

It is important to realize that the error rate will remain at 5% if the experimenter only monitors one time when traditional A/B testing is used. For every additional monitoring period, 5% will be added to the error rate. For example, taking 4 “peeks” will result in an error rate of 20%. However, with sequential testing the error rate will remain below 5% even after 4 “peeks”.

Another big improvement Optimizely made was switching from controlling the Type 1-error rate (or false positive rate) to controlling FDR. In traditional statistics A/B testing methods, using Type 1-error rate control, testing multiple goals and variation at once can introduce problems such as “the multiple testing problem”.

Looking at Figure 3-2, consider testing 5 variations of your website, each having 2 goals. One of the variations positively outperforms and is correctly declared a winner. By having a statistical significance level of 90%, we would expect about 1 more variation falsely declared a winner (10% of the other goal-variations combinations). We now have 2 variations declared winners, although we controlled for a 10% false positive rate (1 false positive). This leads to a 50% chance of making an incorrect business decision. This 50% is also called FDR and Optimizely reports winners and losers with low FDR rather than a low false positive rate.

An experiment with ~ 10% false positive rate and 50% false discovery rate

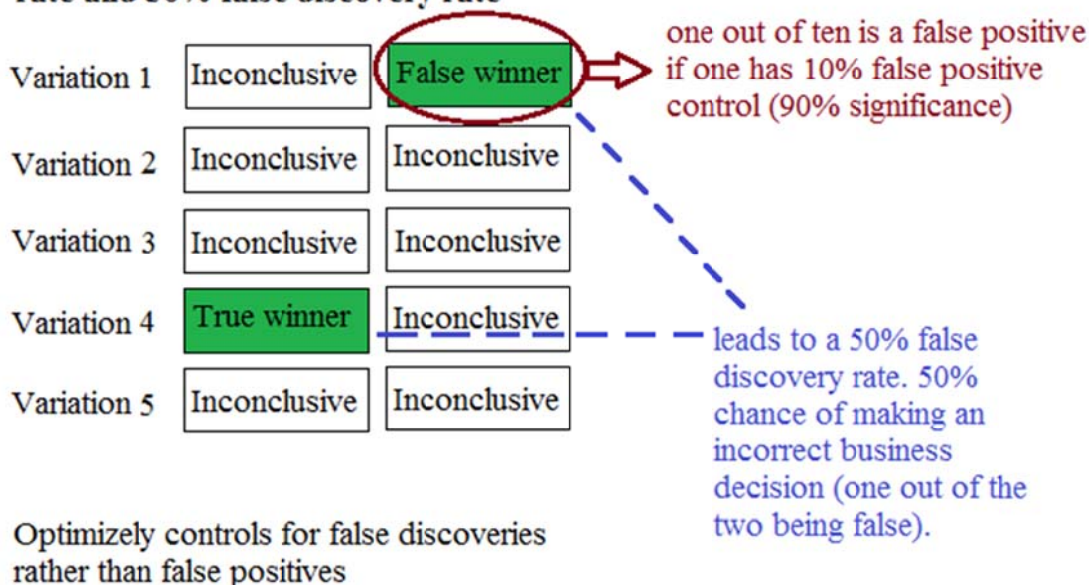


Figure 3-2: Multiple testing problem

FDR control is designed to control the expected proportion of “discoveries” (rejected null hypothesis) that are false (incorrect rejections). As FDR is defined as the expected proportion of false detections among all detections made, or in mathematical terms: $FDR = E[V/(RV_1)]$, in which V is the number of significantly declared tests which are truly null and R counts the overall number of tests declared significant. It has been shown by Benjamini and Hochberg that any hypothesis testing procedure designed to control Type 1 error rate can be transformed into one for controlling FDR using a method [35]. This method underlies Optimizely’s multiple testing approach [33].

3.5 Planned Data Analysis

The subsequent sections will describe the data analysis technique and the software tools in use.

3.5.1 Data Analysis Technique

We will use Optimizely’s Stats Engine to determine significant results as described in Section 3.4.8. However, we will compare these results with a traditional significance test as described in Appendix A.

3.5.2 Software Tools

The software tools used will be Optimizely and Google Analytics.

3.6 Evaluation framework

As mentioned in Section 3.1, the framework we used added additional procedures to an original A/B testing framework, meaning that it can only be seen as an improvement of the current up-to-date approach. Section 3.4.1 argued for why the choice fell for a quantitative approach with supporting qualitative research.

4 Results

This chapter will first present Google Analytics findings related to developing personas described in the research process step 1.a outlined in Section 3.1. Next we present Google Analytics findings relating to gaining insight into where the optimization process should start, described in research process step 1.b outlined in Section 3.1. After that the design proposals will be presented. Lastly, the results of the A/B tests will be presented.

4.1 Extracted Google Analytics User Data for Personas

The results of the Google Analytics findings have been accessed through the reporting's navigation panel. We will describe the steps taken to locate the desired information with the use of arrows (→), as in Audience → Demographics → Overview which describes first clicking the Audience tab, then the Demographics tabs, and finally the Overview tab. The time span of the data collection is 29 April 2015 to 1 April 2016 (11 months). April 29th 2015 was the day Nordic Design Collective filtered out all the bots from analytics sessions, making data more reliable from that day onwards.

The results of step 1.a: Getting to know the website's main users are described in the following subsections.

4.1.1 Step 1: Audience → Demographics → Overview

Figure 4-1 shows that 80.8% of all sessions to the website are by women, while 19.2% are of men, a 4:1 ratio.

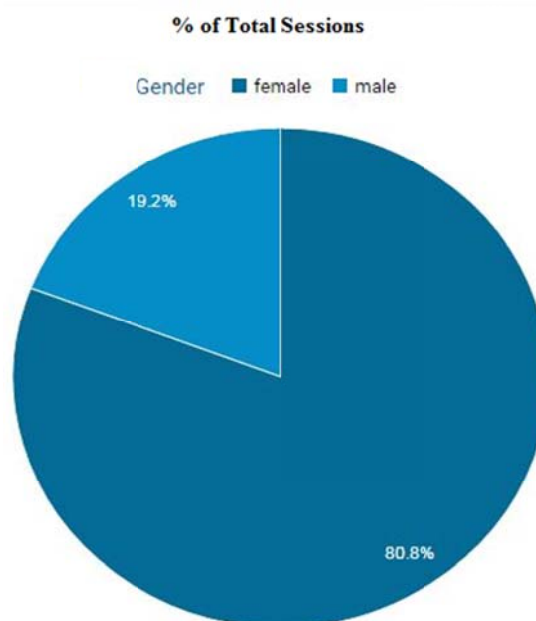


Figure 4-1: Session distribution among males and females

4.1.2 Step 2: Audience → Demographics → Gender → Secondary dimension = Age

From Figure 4-2, we can see that the largest group of people visiting the website is women between the ages 25-34. The second largest group is females aged 35-44. While considering males, the age between 25-34 and 35-44 are also the largest groups. This step helps us realize what gender and age group(s) we should design for or at least those combinations we should consider when making a design decision. This knowledge is important because different design styles would be made to adapt the site to different age groups.

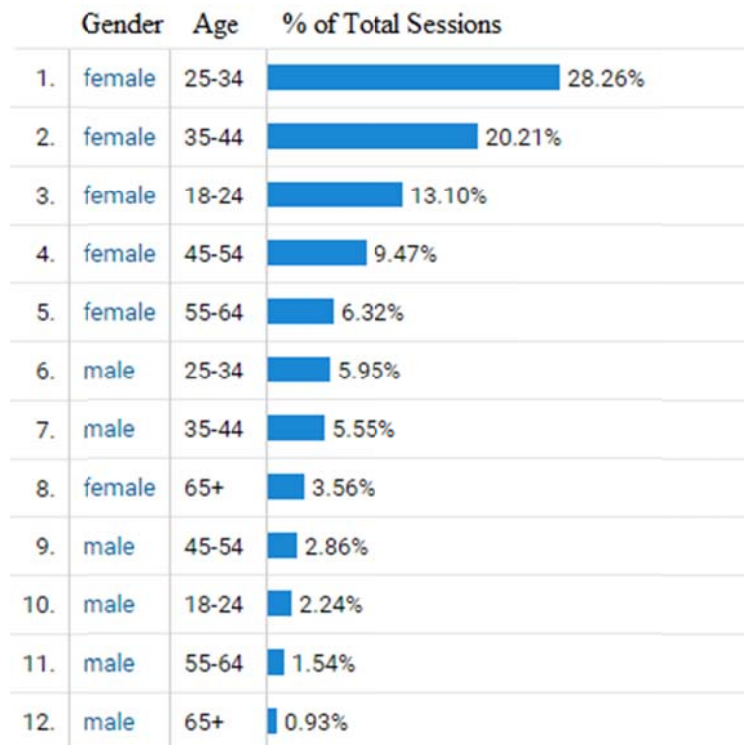


Figure 4-2: Percentage of sessions of various target groups

4.1.3 Step 3: Audience → Interests → Overview → Set Segment: Males, Segment: Females

The shared affinity or interests amongst males and females ordered from most popular to least is shown in Figure 4-3. Men who visit the website are mostly Movie & TV lovers, technophiles, and/or have an interest in online videos and sports. Women, on the other hand, have an affinity for home decor and are cooking enthusiasts; some have interests in celebrities and entertainment news as well as food, drink, and recipes. Both genders have a high interest in Arts & Entertainment. The interests' category allows us to design in a category-specific way based on users' likes rather than the designer's personal taste.



Figure 4-3: An overview of male and female interests

4.1.4 Step 4: Audience → Mobile → Devices

Figure 4-4 shows that half of the people browsing through the site using a mobile device use an Apple iPhone, while a quarter of them use an Apple iPad. Third on the list, with only a couple of percent of the visitors are visits through Samsung Galaxy devices. The importance of which device is used suggests that we need to consider the responsiveness of the website’s design (and our proposed changes to the design) on Apple devices as they are used most. Of course, the responsiveness should be equal on all devices, but strongest focus should be placed on the Apple devices as this will have the greatest effect upon the actual visitors to the site.

	Mobile Device	% of Total Device Sessions
1.	Apple iPhone	50.74%
2.	Apple iPad	24.49%
3.	Samsung SM-G900F Galaxy S5	2.84%
4.	Samsung SM-G920F Galaxy S6	2.17%
5.	(not set)	1.19%
6.	Sony D5803 Xperia Z3 Compact	1.13%
7.	Samsung SM-G925F Galaxy S6 Edge	0.87%
8.	Samsung I9506 Galaxy S4	0.68%
9.	Sony D6603 Xperia Z3	0.55%
10.	Sony D5503 Xperia Z1 Compact	0.54%

Figure 4-4: Within mobile device usages, the percentage of sessions made through specific devices

4.1.5 Step 5: Customization → Hourly & Daily Engagement (Hour, Date & Day)

The hourly engagement, displayed in Figure 4-5, allows us to visualize in our scenario not only when the personas are most likely to visit the website, but also what they have done recently and what their moods could be. In our case, increases in visits occur around 8 AM, which might lead the designer to believe that these visitors are visiting the page on their way to work, perhaps on public transport. Another increase occurs after 8 PM, typically after people have had dinner and are resting & relaxing.

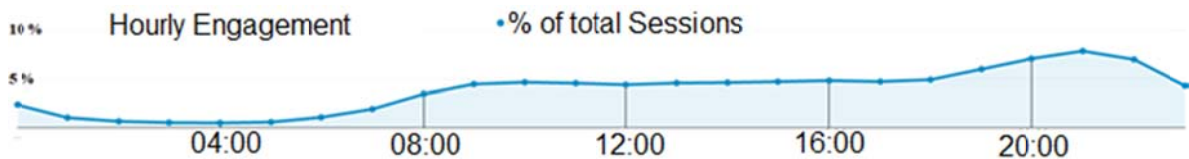


Figure 4-5: Percentage of sessions hour by hour (hourly engagement on the website).

4.1.6 Step 6: Audience → Geo → Location

Visitors from the Scandinavian countries listed in Figure 4-6 together make up 85.58% of all sessions, with visitors from Sweden comprising 78.85% of all sessions.











	Country	% of Total Sessions
1.	 Sweden	78.85%
2.	 Finland	3.68%
3.	 United States	2.96%
4.	 United Kingdom	1.98%
5.	 Norway	1.92%
6.	 Denmark	1.48%
7.	 Germany	1.01%
8.	 Australia	0.69%
9.	 France	0.66%
10.	 Netherlands	0.63%

Figure 4-6: Distribution of sessions amongst countries

4.2 Persona One – Karin

Our first persona created from the data collected is shown Figure 4-7, namely, Karin. Karin is a 31-year-old social worker who has an interest in artwork and foods. She lives in Stockholm, Sweden, together with her partner and 6-year-old daughter Emma. Karin loves decorating her home, so on her way to work she visits home decor websites through her iPhone 5s and browses through various products. She does the same thing later after dinner when she is at home, but now she finalizes the purchase through her iPad. Nothing pleases Karin more than coming home to her partner and Emma, snuggling under the blankets and watching a romantic comedy.



Figure 4-7: Persona of the main target group, Karin, 31.

4.2.1 Scenario - Karin

Karin is on her way to work and has a 30-minute commute by train until she begins to walk again. During this half hour, she listens to music and starts thinking whether there is something out there that could improve the beauty of her home (and possibly make her stand-out whenever her friends visit). She lands on Nordic Design Collectives homepage and does not know what she wants. She scrolls 3-5 times before she looks at the menu and asks herself what she wants as she looks through the categories of products.

4.2.2 Scenario Insights - Karin

Karin's scenario suggests a design in which appropriate product suggestions should be available to inspire her. What kinds of suggestions are made to her while she is scrolling down the site's home page? Are there enough categories of inspiration within 3-5 scrolls? Is there an inspiration category implemented to further help someone who does not know what they seek? In this scenario the focus lies in browsing.

4.2.3 Persona Two – Stefan

Our second persona is shown in Figure 4-8, Stefan, the 28-year-old "tech-junkie" who works with web development. Being in front of a computer all day allows him to take minor breaks and browse the Internet whenever he wants. He lives in

Stockholm and is single and ready to mingle. His biggest interests apart from watching cute cat videos on YouTube are to read articles about tech inventions and to watch movies. He loves every movie made by Christopher Nolan and is a huge fan of the Batman franchise.



Stefan
"The tech-junkie"

information usage
info Posters, Artwork
%Mobile 30
Mob loc work, commute

"There are 10 types
of people, those who
understand binary,
and those who don't"

Demographics

age 28
occupation Web Developer
location Stockholm
marital status Single
children No
net income 34 500kr/month
education IT Graduate
hobbies Video & Movie lover,
Technophile, Arts & Entertainment

Device usage

computer Macbook Pro
cell phone Iphone 5s
other
primary device Laptop
comfort 10-17
web 8 hours/day
phone 2 hours/day
programs Web dev tools, Facebook, Google, Youtube

Figure 4-8: Persona of the main target group – Stefan, 28

4.2.4 Scenario - Stefan

Stefan just came to work and is taking his first sip of coffee of the day. His co-worker Daniel pops into to his cubicle and discusses the superhero movie he watched last night. During the discussion, Stefan justifies why that superhero is not the best one. After the discussion when Daniel has left, Stefan seeks some artwork or poster with his favorite superhero, Batman, in it. He lands at Nordic Design Collective's homepage and immediately navigates to the search bar, knowing what he wants. He types in "Batman" and waits to see what happens. If nothing comes up, he heads to the "Tavlor & Posters" category and seeks a subcategory related to superheroes. If he finds what he seeks, he immediately wants to know how much the total cost will be including shipping, as he does not want to get his hopes up for something that will end up costing more than expected.

4.2.5 Scenario Insights - Stefan

Stefan's scenario suggests a design in which the search tools are effective and information about the products total price are clearly presented. The subcategories within categories should also be intelligently categorized.

4.3 Finding where to Start - Inspect highly valued and trafficked pages

The potential pages to optimize were those pages that met the criteria of (1) Having made sales over a threshold of X SEK* and (2) Having an exit-rate of over 25%. Our first step is to investigate the sales performance of all products pages by going through Conversions → E-commerce → Product Performance in the navigation panel. Then, sorting the list by Product Revenue (descending order) will allow us to see which product pages bring the most revenue. We have to make a note of these pages because the next step is to paste each page one by one into the Exit Pages report which can be inspected through Behavior → Site Content → Exit Pages. After pasting each page into the search box we can then inspect each page one and take note the value in the %Exit column.

Potential Exit rate pages are shown in Table 4-1.

Table 4-1: High-valued pages


Page	Criteria 1: Revenue	Unique purchases	Page Views	Criteria 2: % Exit- rate
Haväng	Yes	X	~1 000	46.24
Kranarna	Yes	X	~1 500	27.36
12 FLOWERS Calendar 2016	Yes	X	~1 200	34.00
Hängare för posters (30x40 cm)	Yes	X	~1 600	32.83


The only pages that simultaneously met criteria one and two in Nordic Design Collective's website had a very small number of page views (visits) in relation to the time span of the collected data. With 1600 page views over 14 months of data, i.e., 114 visitors per month, this is an incredibly low number of visitors, far below significant in an A/B testing time plan. For this reason, we chose to make a design change on *all* product pages during our first A/B test.


4.4 A/B testing Design Proposal 1 – Before and after

Figure 4-9 displays the original design of a product page in Nordic Design Collective's website. Figure 4-10 displays the experimental design variation.

* Note that the threshold price in Swedish kronor and the actual number of unique purchases in Table 4-1 have been replaced by "X" as this data is proprietary to the company.


FRI FRAKT VID KÖP ÖVER 400 KR | SÄKRA BETALNINGAR | VÅR KUNDTJÄNST | 

LOGGA IN
VARUKORG 

HANDLA HÄR
NORDIC DESIGN COLLECTIVE
INSPIRATION
Sök här 

JUL | INREDNING | KÖK | TAVLOR & POSTERS | PAPPER & KONTOR | SMYCKEN | ACCESSOARER | BARN
VISA FORMGIVARE ▾

TAVLOR & POSTERS • FOTOKONST • HAVÄNG
NEW ARRIVALS | DAN ISAAC WALLIN | INSPIRERAT AV NATUREN | AFFISCHER | PLANSCHER | KONST | TENTI LONDON 2016 | SVARTVITA POSTERS



DAN ISAAC WALLIN HAVÄNG

HAVÄNG

BY: DAN ISAAC WALLIN

Fotokonst poster tagen med polaroid kamera. Ingen digital bildbehandling.
Bilden är från Haväng, Österlen, Sverige.

450 KR

VÄLJ RAMSTORLEK ▾

MEDELANDE TILL FORMGIVAREN VID KÖP

Utskrivet på 200g papper

Bildbeskrivning:
1. Postern du kommer att få.
2. Hur det kan se ut på vägg.
3. Posters

LÄGG I VARUKORGEN

ANTAL:

Tack för att ni tittat på mitt arbete!

Leveranstid: 5-7 vardagar
Lagerstatus: I lager

Kontakta gärna vår kundtjänst om du har några frågor!

Figure 4-9: Original design of a product page

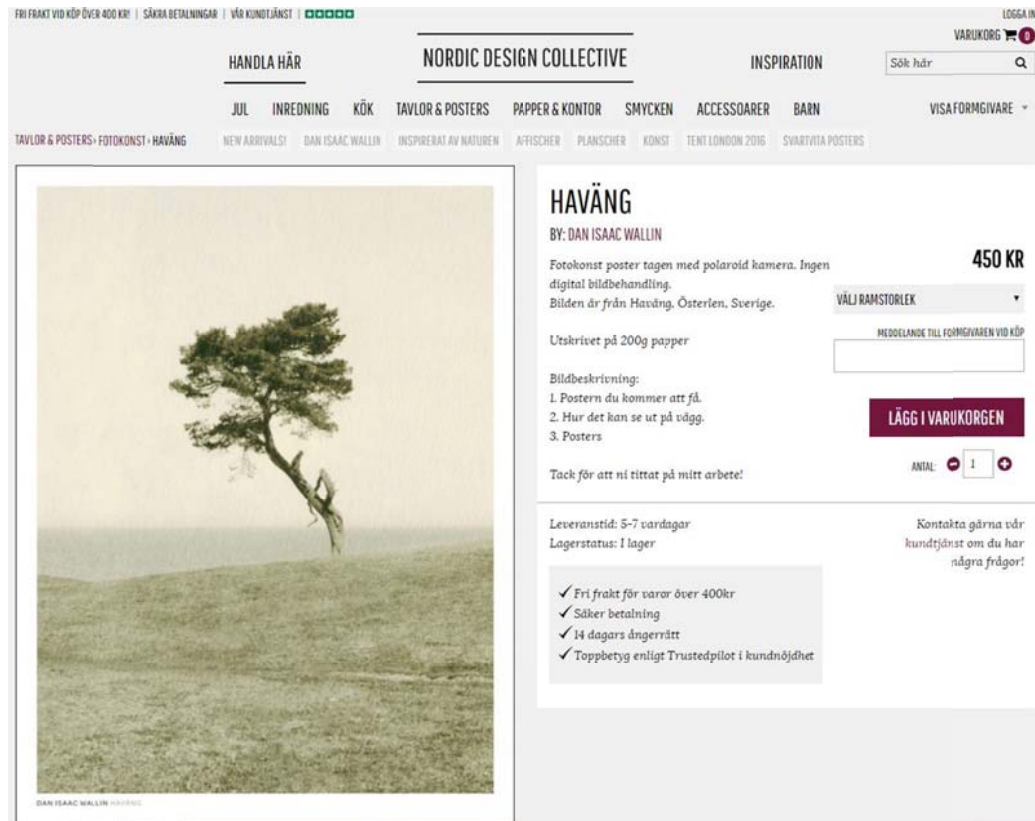


Figure 4-10: Experimental design of a product page

As said in Section 2.2, the process of conversion optimization involves making small changes in an iterative manner to identify whether design changes bring success or not. The drawback of making major visual changes on only parts of a website with a design that diverts from the original template is that it will make the website look unprofessional, and in turn, untrustworthy. For this reason, the focus is less on the visual design and more on designing solutions to important factors for users based upon the findings in Sections 2.10 and 2.8 and to try improving the functionality by considering the design principles given in Table 2-1. This approach seemed best suited since Nordic Design Collective were not looking to improve the conversion rate of *only a single* target group, but rather *all* their target groups. Their goal was to increase the conversion rate for all target groups, with a general design solution.

From the qualitative study shown in Table 2-2 - regarding what users look for in e-commerce stores web layout when deciding to shop, the main important features other than clear information and the total price being presented, are that the site providing evidence of security and instils trust. This is what we attempted in our design of the information box shown in Figure 4-10. Our design hypothesis thus became:

If we add a prominent information box that instils trust and encourages shopping through the promise of free delivery, users will feel more comfortable and willing to make a purchase.

Users tend to look for a secure e-commerce certificate and/or customer reviews to gain trust in a website where they are considering making purchasing. Since Nordic Design Collective did not have a certificate, but has a secure payment method, as well as outstanding customer reviews (which are visible on another part of the site), we thought to promote the safe payment method, the top customer review scores, and to include text about the 14 days return guarantee of to instill trust and assurance of quality. Nordic Design Collective does a good job in displaying the total price as early as possible when looking at the cart, but we thought to make this somewhat clearer earlier than that by adding a bullet indicating free shipment for products over 400 SEK. This allows the user to directly do the math and know if there will be any hidden costs or not as soon as they access a product page. This information box element was strategically placed close to the “LÄGG I VARUKORGEN” (place in shopping cart) button and purposefully has a grey background to make it stand out, but not depart too much from the original design. The check-mark symbols bullet-points were added to attract the user’s attention.

There is also a functional improvement made if one pays close attention to the amount (“antal”) box beneath the “LÄGG I VARUKORGEN” button. The new design has two buttons in form of a plus (+) and minus (-) that incrementally adds or removes items if one chooses to click on them. Looking at Figure 4-9, the design does not give enough clues as to how to use the amount box. For instance, from Figure 4-2, we saw that 20.89% of all sessions to the site are people above the age of 44, a generation perhaps not having the same mental computer interaction model as younger users. This functionality was added due to principle 6.2 in Table 2-1, concerning Tesler’s law of the conservation of complexity. We felt the end user’s interaction would be further simplified. We ensure that the design would be appropriate by following the principles in Table 2-1. Principle 2.1 is satisfied as the coloring of the plus/minus buttons are the same as the “LÄGG I VARUKORGEN” button, indicating that it operates as a button as well. Principle 2.2 is satisfied as the symbols plus and minus inform the user in a very well-known way that something is either being added or subtracted, which in this case is the quantity of the item to be purchased. The feedback principle 4 is satisfied as the user receives instant acknowledgement from the system as the quantity changes after a click on either button. The Poka-Yoke principle 3.1 is also satisfied as the function does not allow the user to reduce the quantity below the value of 1, ensuring proper conditions exist before order processing begins and preventing problems from occurring in the first place. Principle 5.1 is also satisfied as the buttons are strategically placed with the minus button on the left side of the box, indicating that clicking on it will lead to a value less than itself (<) and likewise having the plus button on the right side of the box indicating a value greater than itself (>).

4.5 Results of A/B test 1: Info-box and plus and minus functions

According to Optimizely's statistical engine, as seen in Figure 4-11, the first A/B test gave inconclusive results as the results were only 61% statistically significant. Our KPI which we counted as a conversion was computed by tracking the "LÄGG I VARUKORGEN" button on all product pages. The original baseline design resulted in 980 unique conversions in 13 980 visits, while our alternate design resulted in 1 078 conversions in 13 992 visits, a difference of 98 additional conversions. The original design had a conversion rate of 7.01% and our design 7.70%, a relative conversion rate improvement of +9.9%. The difference is used to calculate the confidence interval of our results in terms of a desired statistical significance level of 95%. To conclude that a design is better than the original with a statistical significance of 95%, the difference interval should be positive (i.e., greater than 0).

Our design has a difference interval of -0.27 to +1.66, meaning that the confidence interval in which we would be 95% certain that the true conversion rate of our design will lie between is a conversion rate interval of = (Original conversion rate + the lower limit difference) to (Original conversion rate + the upper limit difference) = (7.01 + (-0.27)) to (7.01 + 1.66) = 6.74% to 8.67%. In other words, we are 95% confident that the design change's true conversion rate will lie between 6.74%-8.67%. Since the lower limit of our confidence interval (6.74%) is less than the original design's baseline conversion rate (7.01%), it is obvious that the design is not positively successful in a statistically significant manner. As this would have required that the lower limit of our design must lie above 7.01%.

When deciding how long the test would run, we made use of the decision tree in fig 6-2 and realized that we could not afford to wait any longer than 70 days to reach significant results.

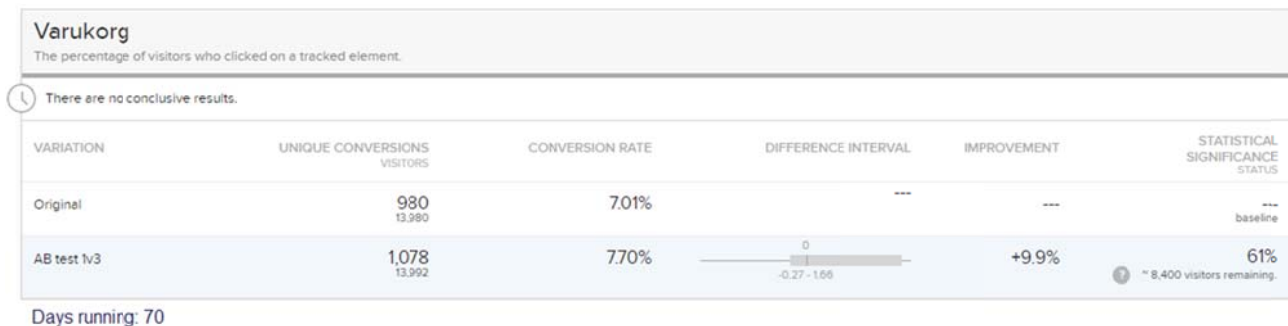


Figure 4-11: A/B test results of Design 1: Info-box and + and - function

4.6 Going Where the Rainbow Ends - Inspect the Check-out Funnel Page(s)

Due to the low significance of the conversion rate difference from the original design, one can either make the changes of the design more drastic or try to optimize something completely different. As Nordic Design Collective wanted a cohesive design which does not deviate radically from its current state, we decided

not to pursue optimizing the product pages and to instead optimize other critical parts of the page. From a sales perspective, the check-out page(s) are considered the most valuable pages since they are pages where users are one click away from making a purchase. Inspecting the check-out funnel reveals valuable exit-rate related information that can be compared to a global average exit-rate percentage. The goal is to reduce the exit-rate of the check-out funnel such that users exit the session via a “Your order has been processed” or “Successful Order” page.

To find this data via Google Analytics we head to Funnel Visualization through Conversions → Goals → Funnel Visualization.

We can see from Figure 4-12 that of 23 243 users who continued to the shopping cart page (Varukorg), 41.17% of them continued to the check-out page. Of the users flowing to the check-out page, 58.50% of them proceeded to make a successful purchase, meaning as *high as 41.50% of visitors abandoned this page and did not go through with a purchase*. Of those who abandoned the final check-out page, 40.63% exited the web page completely (1 600/3 972) while 28.15% returned back to the cart and abandoned it later (1 118/3 972). This means that *as much as 31.22% (100% - (40.63% + 28.15%)) of those who abandoned the check-out page used the navigation menu*. For this reason, our alternative design proposal became that shown in Figure 4-14 (compare with the original design shown in Figure 4-13).

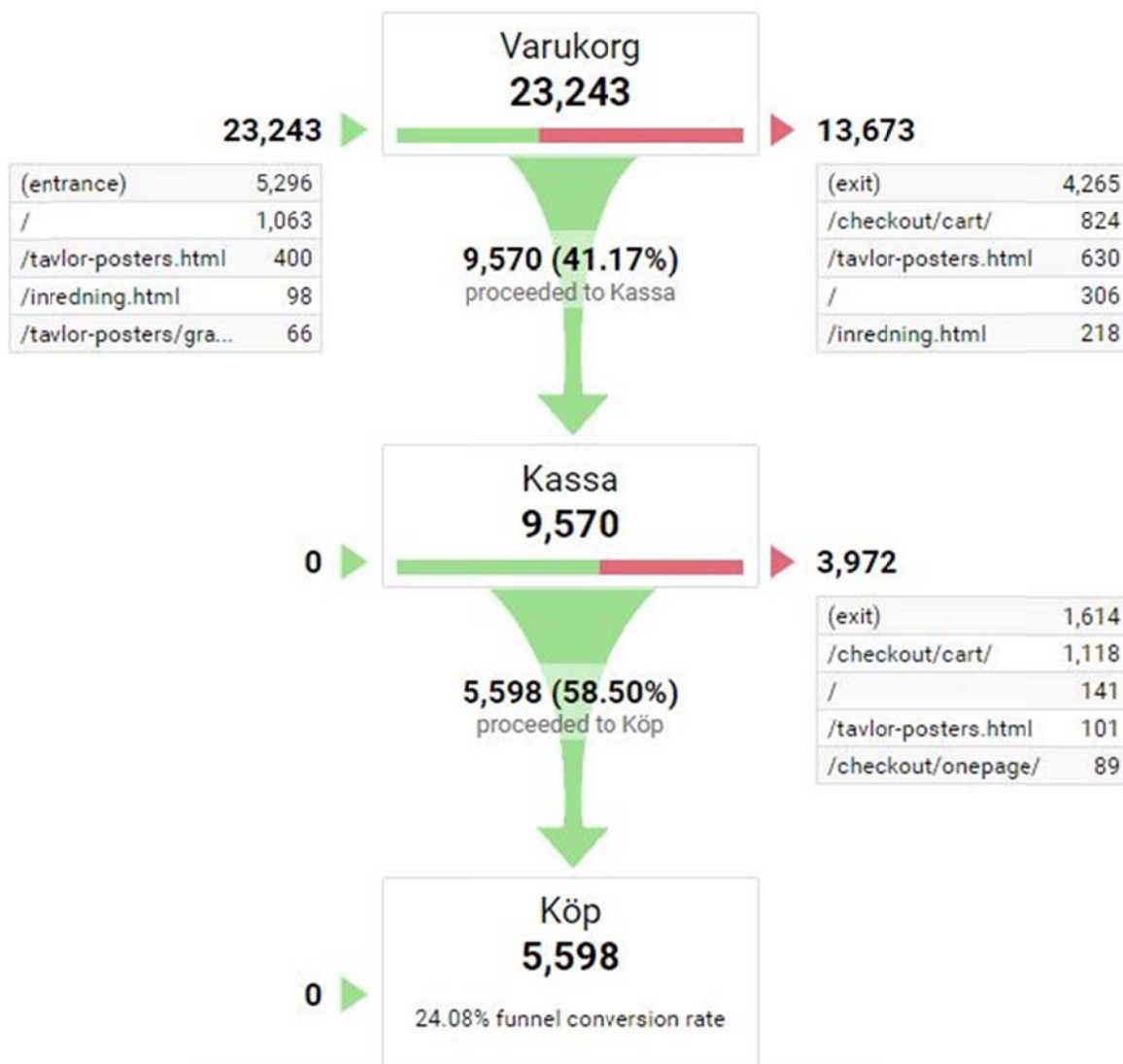


Figure 4-12: Funnel visualization of sessions flowing from the cart page to completing a purchase.

FRI FRAKT VID KÖP ÖVER 400 KR! | SÄKRA BETALNINGAR | VÅR KUNDTJÄNST | 000000

LOGGA IN
VARUKORG 1

HANDLA HÄR
NORDIC DESIGN COLLECTIVE
INSPIRATION

Sök här

JUL | INREDNING | KÖK | TAVLOR & POSTERS | PAPPER & KONTOR | SMYCKEN | ACCESSOARER | BARN
VISA FORMGIVARE

KASSA

Välkommen till kassan. Fyll i fälten nedanför för att slutföra din beställning!

Har du redan ett konto? [Klicka här för att logga in.](#)

1. FAKTURAADDRESS

PERSONNR./ORGNR

FÖRNAMN * EFTERNAMN *

E-POSTADRESS * TELEFON *

ADRESS *

LAND *
SVERIGE

STAD *

POSTNUMMER *


FÖRETAG



LEVERERA TILL SAMMA ADRESS


2. LEVERANSSÄTT

FRI FRAKT
0 KR

3. BETALSÄTT

FAKTURA (29 KR) 
VILLKÖR FÖR FAKTURA

KORTBETALNING (0 KR)  

PAYPAL (0 KR) 

4. KONTROLLERA DIN ORDER

Produkt	Antal	Summa
Mira Necklace SG	1	990 kr
Summa		990 kr
Faktureringsavgift		29 kr
Frakt (Fri frakt)		0 kr
Totalt		1 019 kr

PRENUMNERA PÅ VÅRAT NYHETSBRÄV

Godkänn våra köpvillkor genom att klicka i rutan här nedanför.

JA, JAG GÖRKÄNNER KÖPVILLKOREN

LÄGG ORDER

Figure 4-13: The original version of the check-out page.

FRI FRAKT VID KÖP ÖVER 400 KR! | SÄKRA BETALNINGAR | VÅR KUNDTJÄNST | 00000

NORDIC DESIGN COLLECTIVE

KASSA

Välkommen till kassan. Fyll i fälten nedanför för att slutföra din beställning!
Fortsätt att handla

Har du redan ett konto? [Klicka här för att logga in.](#)

1. FAKTURAADDRESS

PERSONNR/ORGNR

FÖRNAMN * EFTERNAMN *

E-POSTADRESS * TELEFON *

ADRESS *

LAND *
SVERIGE

STAD *

POSTNUMMER *



FÖRETAG



LEVERERA TILL SAMMA ADRESS


2. LEVERANSSÄTT

FRI FRAKT
 0 KR

3. BETALSÄTT

FAKTURA (29 KR) 

VILLKÖR FÖR FAKTURA

KORTBETALNING (0 KR)  

PAYPAL (0 KR) 

4. KONTROLLERA DIN ORDER

Produkt	Antal	Summa
Mira Necklace 3G	1	990 kr
Summa		990 kr
Faktureringsavgift		29 kr
Frakt (Fri frakt)		0 kr
Totalt		1 019 kr

PRENUMERERA PÅ VÅRAT NYHETSBRÉV

Godkänn våra köpvillkor genom att klicka i rutan här nedanför.

JA, JAG GODKÄNNER KÖPVILLKÖREN

LÄGG ORDER

Figure 4-14: The experimental design variation of the check-out page.

Nordic Design Collective's check-out page is very effective in its design, as all the steps needed to complete a purchase are clearly ordered and placed on a single page. The design is cleverly ordered horizontally as this allows the user to avoid muscular tension (Table 2-1, principle 5.2) since there is no scrolling movement required (as would have been the case in a vertical design). The total price and methods of payment are clearly presented and there is no forced registration.

As mentioned earlier, 31.22% of those who abandoned the check-out page and never went through with a purchase abandoned the check-out process by using the navigation menu. As seen in our design proposal shown in Figure 4-14, many navigation elements are removed when compared to the original version displayed in Figure 4-13. Our design hypothesis was:

If we remove re-directing navigation elements, users will be less distracted and have less incentive to back out of a purchase commitment and thus are more likely to make a purchase.

The navigation bar is of course of great importance, as it allows users to get inspired and seek what they are interested in. However, it seemed counter-intuitive to have an option that *encourages* users to continue browsing on such a critical page. From a sales perspective, we do wish not to invoke hesitation, but rather we wish to seal the deal as quickly as possible. This is in agreement with Hick's Law (Principle 6.2 Table 2-1) and was the reasoning behind removing the navigation elements.

Our check-out page's reduced options of movement must be compensated to satisfy Table 2-1, entries 1 and 2, by considering how users will interact with this new design if they wish to go to the homepage to continue shopping. To address this, a new design with a purple link with the text "Fortsätt Handla" ("Continue Shopping") has been placed below the welcoming check-out text. This link, which on hovering changes color, satisfies principles 2.1 and 2.2 in Table 2-1. The link is the same color as other links on the site giving the user a clue that clicking it will redirect the user back to the homepage (providing affordance). The change in color on hover provides the user with information about what will happen if one chooses to click it (providing feedforward).

4.7 Results of A/B test 2: Removing Check-out Navigation

According to the Optimizely results, shown in Figure 4-15, the A/B test gave *inconclusive* results as they were not statistically significant. This would not have been the case if one used traditional fixed-horizon statistics as the variant would have (*incorrectly*) been declared a statistically significant winner. Our KPI which we counted as a conversion was done by tracking the place order "LÄGG ORDER" button. The original design resulted in 370 unique conversions in 553 visits, while our alternative resulted in 445 unique conversions in 613 visits, a difference of 75 additional conversions. However, unlike our first A/B test when comparing the number of visitors to each design we can calculate that the original received 47.43% of visitor flow, while the alternative received 52.57%. Although we had pre-set the visitor flow to direct 50% of visitors to each design, Optimizely had not yet balanced the flow equally at the time we chose to end the test.

When deciding how long the test would run, we made use of the decision tree in Figure 6-1 and realized that we could not afford to wait any longer than 60 days to reach significant results.

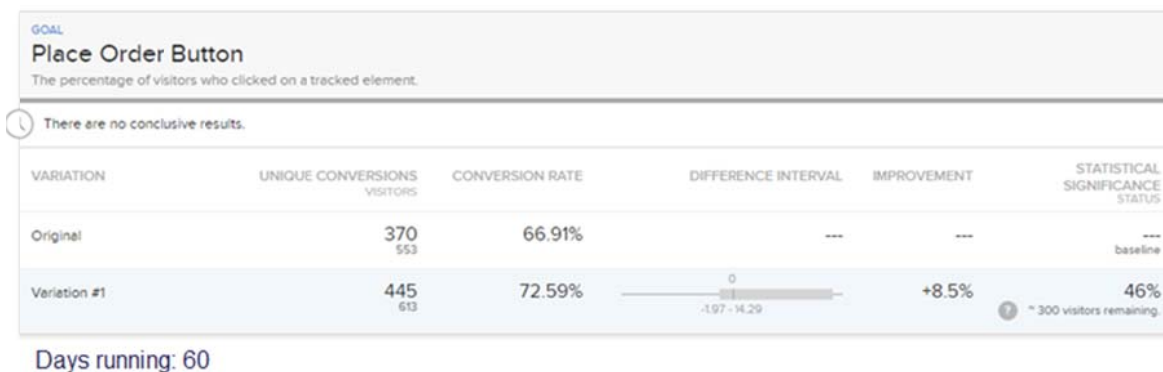


Figure 4-15: A/B test results of Design 2: Removing Check-out Navigation.

The original design had a conversion rate of 66.91% and our design had a conversion rate of 72.59%, giving a conversion rate improvement of +8.4%. The difference interval of -1.97 to 14.29 estimates our design's confidence interval of our conversion rate ranging somewhere between 64.94% - 81.20%. As the lower limit of our confidence interval falls below the original design's baseline conversion rate, we cannot with 95% confidence say that our design performs better. The lower limit of our variation's confidence interval would have to be over 66.91% in order for it to be declared a statistically significant winner.

5 Analysis

In this chapter we will discuss, in Section 5.1 our analysis regarding the major results of our A/B tests, what business decisions can be made from them and the risks. Section 5.2 presents a reliability analysis, while Section 5.3 presents a validity analysis. Section 5.4 discusses the challenges we faced, the time required to reach significance, and some suggestions to others working in this area and what we would have done differently if we were to do this work again (see also Sections 6.2 and 6.3 in the following chapter).

5.1 Major results

Section 5.1.1 starts off introducing how one should interpret Optimizely's results to make business decisions. Sections 5.1.2-5.1.3 discuss the results of both A/B tests.

5.1.1 Interpreting Optimizely's Statistical Engine to make business decisions

As both of the A/B tests gave inconclusive results, the question of whether implementing the variations would be worthwhile or not arises. In addition, there are the questions of: Can we make any business decisions based on inconclusive results, specifically on results that are not statistically significant? To answer this question, we need to discuss the difference intervals of our results.

The difference intervals inform us of the range of values where the difference between the original and alternative actually lies, after removing fluctuation. It is a confidence interval of the conversion rates that one can expect to see if one chooses to implement the alternative in question. We can consider this interval as the "margin of error" in the absolute difference between the two competing conversion rates. Optimizely's difference interval should either lie entirely above (winning variation) or below (losing variation) 0% if there is a statistical significance in the two version of the page (or site). Conversely, inconclusive results, such as ours, have a difference interval which includes 0%.

In analyzing the results of our A/B tests, we must keep in mind that Optimizely set our difference interval at the same level as our 95% statistical significance threshold for the project.

5.1.2 A/B Test 1

The result of our first A/B test, shown in Figure 4-11, says that the difference in conversion rates for this design will be between -0.27% and 1.66%, meaning that it could be positive or negative. The original baseline conversion rate is 7.01% and if we want to make a business decision about whether implementing this variation will be worthwhile; we can put it in terms of worst case/middle ground/ best case scenarios.

We are 95% confident that the worst case absolute difference between the variation and baseline conversion rate is -0.27%, the best case is 1.66%, and a middle ground (average) is 1.39%. This means that if we were to implement this design, the worst case scenario is a 0.27% decrease in conversion rate, the best case scenario is a 1.66% increase in conversion rate, and an average case scenario is a 1.39% increase in conversion rate.

This being said, although the results never reached significance, we would recommend Nordic Design Collective to implement this design since the risk is less than the reward with an average case scenario of 1.39% conversion rate increase. If Nordic Design Collective wants to lower their risk of conversion loss, they could continue to let the A/B test run while setting the desired percentage of visitor flow of as low as they want.

5.1.3 A/B Test 2

The result of our second A/B test, shown in Figure 4-15, says that the difference in conversion rates for this design will be between -1.97% and 14.29%, meaning that it could be positive or negative. The original baseline conversion rate is 66.91% and if we want to make a business decision about whether implementing this variation will be worthwhile; we can put it in terms of worst case/middle ground/best case scenarios.

We are 95% confident that the worst case absolute difference between the variation and baseline conversion rate is -1.97%, the best case is 14.29%, and a middle ground (average) is 6.16%. This means that if we implement this design, the worst case scenario is a 1.97% decrease in conversion rate, the best case scenario is a 14.29% increase in conversion rate, and an average case scenario is a 6.16% increase in conversion rate.

This being said, although the results never reached significance, we would recommend Nordic Design Collective to implement this design since the risk is less than the reward with an average case scenario of 6.16% conversion rate increase. If Nordic Design Collective wants to lower their risk of conversion loss, they could continue to let the A/B test run while setting the desired percentage of visitor flow of as low as they want.

5.2 Reliability Analysis

We have now seen the hard data saying that our A/B tests made an improvement concerning conversion rate. How can we trust this data? One of the points we mentioned in Section 2.12 is sampling -- did this occur? When Optimizely was running the A/B tests, there was no sampling occurring that might have affected the results. Nor was there any sampling in the Google Analytics data. However, there could be incomplete data if all pages meant to run the A/B tests did not actually do so. In order to ensure that we ran the first experiment on all pages that

we were supposed to, we chose to include the experiment all pages, but explicitly excluded those specific pages that were not product pages. This was done by an option in Optimizely called Page Targeting. To know which specific pages were not product pages we manually visited all pages on the website that were not product pages and pasted the Uniform Resource Locator (URL) for these pages into Optimizely. We then verified with the website manager whether we had missed anything. For the second experiment, incomplete data was not an issue as the experiment was run only on the checkout page.

Another factor to consider is if the experiment actually evenly distributed visitors between the A/B versions. Figure 4-11 shows that this was quite true for the first A/B test as there was a 0.09% difference in the distribution of visitors between the two versions. However, in the second A/B test the difference was 5.15%. This could be due to the low total number of visitors. One visitor has a larger impact on a smaller pool and revisiting will not change which version of the website a visitor views, therefore Optimizely cannot effectively balance the visitors over the two versions.

There are other factors that are worth noting, some of these were mentioned in Section 3.4.6. To consider how users behave with regard to deletion or blocking of cookies would require some additional qualitative research to be performed to accurately appreciate its impact on the reliability of the data. To gain some perspective, we looked earlier related research and found that there exist a few studies that attempted to understand individuals' behavior. The largest recent study we found was made by comScore in 2007 which monitored approximately 400,000 home computers during the entirety of the month of December in 2006 [37]. The study found that first-party cookies were cleared within a month by 31% of personal computer users in the United States of America. Other major prior research also mentioned in the comScore study, concluded that at least 30% of users deleted cookies during a month.

These figures might seem out of date, but a more recent study released in January 2011 with a focus on Australia delivered similar results [38]. This later study mentions that roughly 12% reject cookies through various methods, e.g. browser settings. Third-party cookies are also of importance for Google Analytics, when trying to identify demographics and interests of users. Studies show that third-party cookies were deleted by 27% of computer users in the U.S. in 2007. It might seem counter-intuitive that third-party cookies were deleted less frequently than first-party ones, as third-party ones have gained a reputation of being more invasive — but this trend did not continue. For example, between 30-40% of computers in the UK, Germany, Australia, and U.S. during 2011 deleted third-party cookies within the month.

If we consider the applicability of this research to our data, the third-party cookies fall away as a definitive factor. The only use of third-party cookies in this project is Google Analytics' use of the renowned DoubleClick add server persistent cookie.

Although the count of unique visitors may be overstated by up to 170%, based upon these earlier studies, the distribution of data will still remain somewhat similar. Essentially, each metric will be exaggerated differently depending on its magnitude; hence highs will potentially be more represented than lows will. As we are mainly interested in the shape or form of the data distribution and less on the specific values in Google Analytics, therefore the effects of not being able to uniquely identify visitors has less effect on the reliability of our study. Understanding how much these results could be exaggerated is outside the scope of our study.

The use of ad-blocking software has grown the last few years. The frequency of its use differs based on content. Globally the use of ad-blocking has reached 198 million users, according to a report by Adobe and PageFair summarized in 2015 [39]. Ad-blockers appear on both desktop computers and mobile devices. Although the user of ad-blockers on mobile devices is not as common as on desktop computers, their use is on the rise. In Sweden, where most of the traffic at Nordic Design Collective originates from, 25.10% of users use some sort of ad blocking tool. This is pertinent for the analytical stage of the data reviewed in Google Analytics, but the effect is negligible when considering the A/B test itself. However, the effect of ad-blockers is of the utmost importance for advertising agencies and businesses dependent on ads. The data gathered through ad elements on the website may be blocked and prevent user identification, leaving only non-personally identifiable information. Fortunately, this is not an issue for Nordic Design Collective users as there are no external advertising elements on the site.

A few other variables that are not easy for us to account for and quantify their magnitude and importance are the fraction of users using multiple devices and the fraction of users sharing a device. This behavior is common. A study made by Gesellschaft für Konsumforschung (GfK) on more than 2,000 people in the UK and the US concluded that more than 60% use at least two devices every day and more than 40% start their browsing on one device and finish it on another [40]. It is reasonable to assume that the situation is similar for users in Sweden and hence for the major users of Nordic Design Collective's website. As seen in Figure 4-4 on devices, besides desktops the major devices are iPhones and iPads. It seems that people have a preference for making purchases on their computers; they feel that their mobile device's security is insufficient. However, it is difficult to know what fraction of people behave in this manner. If we look at how many people are on the website at specific times and split them additionally into device types we can get a better idea of these users' behavior. The data shows that between 8-10PM 12.19% of all users browse using a mobile device, while 4.20% use a tablet – this means that the vast majority of users are using a desktop computer to browse the website during these hours.

5.3 Validity Analysis

Throughout our research on how to evaluate A/B test results, there has not been a clear-cut, easy to follow, and universally agreed upon method for evaluating the significance of test data. The commonly used methods are a Chi-squared test, t-tests, calculating statistical significance using binomial distributions or Bernoulli, normal approximations of binomial distributions, Fisher's exact test, z-tests, etc. Some of these methods are preferred when evaluating certain metrics, but it is not clear what method to use when evaluating conversion rates. There are multiple online tools for calculating the significance level for A/B test results, but some do not disclose how they are calculated and many reach different conclusions (for the same data). As we evaluated our data using multiple different methods, it became more and more apparent to us that Optimizely's conclusion is questionable, as multiple methods gave us a significance level of above 95% that our result for our second test had a positive effect on the checkout page.

To validate the results given by Optimizely, we would like to recreate the calculations by hand. However, there are far too many unknown values of variables defined in their stats engine report—such as FDR, type 1 error produced by continuous monitoring (checking results prior to ending the test), and decision boundary – making our manual calculation impossible [33]. An alternative approach to compare the validity of Optimizely's results is through traditional frequentist statistical significance calculations and to see how much the compared results deviate. In the following paragraphs we will validate the results using Null Hypothesis Testing Using Z-statistics.

We calculate significance by performing a z-statistical hypothesis test. It is important to remember that we will be comparing two conversion rates (means of random variables), and not actual conversion counts. Since what we are testing are basically Bernoulli trials, either a success (conversion) or failure (non-conversion), the trials will follow a binomial distribution which can then be approximated by a normal distribution under the following conditions, which are in accordance with the central limit theorem: $Bin(n, p) \sim N(np, \sqrt{np(1-p)})$ if $np(1-p) \geq 10$, where n corresponds to the number of visitors and p to the conversion rate of the specified page.

We want to see if the difference in conversion rate is statistically significant. Using our calculations (shown in Appendix A), we attained the results shown in Table 5-1.

Table 5-1: Key z-statistics values for A/B test one and two

	A/B test 1 – Product Page	A/B test 2 – Check-out Page
\bar{D}	0.005067	0.0596
N	70	60
σ	0.022065	0.1797
$\sigma_{\bar{D}}$	0.002637	0.0232
	$Z = \frac{0.005067 - 0}{0.002637} = 1.9215$	$Z = \frac{0.0596 - 0}{0.0232} = 2.569$

Note that N, in this case, corresponds to the total number of days the tests were running (see Appendix A for clarification). Our first A/B test failed to be statistically significant with 95% certainty as our z-score is beneath the 1.96 threshold for a two-tailed test. However, it did exceed the 90% threshold (1.6449). Comparing this to Optimizely’s statistical significance status of 61% shown in Figure 4-11, we can see a distinct difference between this and the aforementioned result.

The second A/B test proved to be statistically significant in a positive manner with more than 95% certainty as our z-score lies above the threshold. Comparing this to the 44% level of significance shown Figure 4-15, we can see that there is an even larger distinction here.

The confidence interval we attained using the z-statistics (shown in Appendix A) for the first A/B test was -0.0466 - 0.0567, while for the second A/B it was 0.0141 - 0.1005. Comparing this to the difference intervals given by Optimizely: -0.00027 - 0.0166 and -0.0197 - 0.1429, we see that there is a noticeable difference.

The values generated from Optimizely’s statistical engine which uses sequential hypothesis testing combined with controlling FDRs for multiple hypothesis testing, compared to the values given by the null hypothesis testing, illustrates the affect sequential testing and FDR-control has on the statistical significance of the results.

5.4 Discussion

Section 5.4.1 discusses the challenges encountered while Section 5.4.2 discusses the “time to reach significance” module in Optimizely. Finally, Section 5.4.3 reviews the ambiguity in Optimizely’s statistical engine.

5.4.1 Challenges due to Poor Integration of Software Tools with One Another

One difficulty experimenters’ face with websites in which the e-merchant does not have total control or access to the website’s entire source code is that this lack of source code access not only limits the experimenter’s ability to fully understanding complex functionality, but also prevents the experimenter from implementing

certain functionality needed to obtain more accurate data. For example, in our case our conversion rate measurement tool (KPI) tracked clicks on buttons: “Add to Cart” or “Place Order”. Although it could be argued that a design with significantly more clicks on such buttons compared to the original is a better design, such a click does not necessarily result in a purchase. Our conversion goals can be considered micro conversions in the sense that increasing their conversion rate will more likely lead to a macro conversion, which is the completing of a purchase. Consider our first A/B test’s “Add to Cart” button; there is not a fixed correlation between the number of clicks and sales that was possible to track. We do not know where the user will subsequently end up, i.e., whether the visitor abandons their cart or completes their purchase. However, we did, try to account for this in our second A/B test as the check-out page was located close to (one click away from) completing a purchase. However, with respect to reliability clicking on the “Place Order” button will count as a conversion, even if the form is submitted with errors.

One way to solve these reliability issues is to implement Optimizely’s revenue tracking goals, something that was not possible for us due to the fact that the e-merchant did not have access to the critical code. E-commerce revenue tracking allows experimenters to track the monetary value of an event (i.e. purchase of a product) and it is used to track how different variations impact total revenue [41].

Our suggested approach to verify how our design proposals impacted revenue was to inspect Google Analytics e-commerce sales performance during the time that the tests were ongoing. Nordic Design Collective has all of its products linked well with Analytics, enabling revenue from pages to be tracked. This can then be used to compare revenues with total number of clicks. However, we failed to keep in mind that both versions of the A/B test were running simultaneously- 50-50% each during that period. This makes it impossible to analyze which version lead to how many purchases and which version brought in how much revenue. Therefore, the experimenter, e-merchant, and the third-party company in control of the website’s source code need to cooperate in order to fully integrate A/B testing tools with revenue- and behavioral analytical tools such as Optimizely and Google Analytics.

5.4.2 Time to reach significance

How long will it take to reach statistical significance? Optimizely says while an experiment is running, if you peek at the ongoing test at a specific time, then the data will become significant within X number of viewers. This remains true if the data distributions remain in the direction they are currently in. This seems fairly reasonable and understandable. However, in the course of our data collection, the number of X viewers did not decrease even though time passed by. The figure for “the remaining number of visitors until reaching statistical significance” stayed the same or grew. The figure should only do so if the conversion rate of the variation decreases relative to the original. Looking at the development of the data one would see that there were fluctuations in the conversion rate of the alternatives, as one

could expect. These fluctuations were nonetheless minor and towards the end usually in the positive direction, yet still the number of “remaining visitors” to significance never diminished. Reading Optimizely’s forum posts regarding this topic, there are multiple testers and users questioning the reliability and accuracy of the number being displayed.

5.4.3 Ambiguity in Stats Engine

Reading the document on Optimizely Stats Engine of kindles many uncertainties. These uncertainties regard the many variables defined by Optimizely that are not really quantifiable, as was mentioned earlier in Section 5.3. One example of such a variable is τ that determines the exact decision boundary [33]. The value of τ is vital for determining the significance and the speed it takes to do so. Optimizely is working with the concept of knowing priors; or previous data-driven assumptions in non-jargonistic terms.

6 Conclusions and Future work

In Section 6.1 we discuss our conclusions regarding our goals, insights gained, suggestions to others working in this area, and what we would have done differently if we were to do this work again. In Section 6.2 we present the limitations that we encountered that restricted our results. Section 6.3 presents some suggestions for future work. Finally, in Section 6.4 some reflections are given.

6.1 Conclusions

We succeeded in improving the conversion rates with both of our designs by using web analytical tools, qualitative studies, and HCI principles during the testing time frames. However, these improvements could be caused by random chance since they were not 95% statistically significant (as assessed using Optimizely). In our second A/B test, we did reach statistically significant results when using the traditional null hypothesis testing approach. However, our first A/B test failed to reach statistically significant results using both approaches.

The most important insight gained, concerns the difficulty in finding a sweet spot between wanting to optimize potential high-valued pages and ensure that the pages reach a certain level of user flow as necessary for an A/B test to be declared scientifically conclusive. Our potential high-valued pages that met certain criteria in terms of the number of visitors, revenue generated, and % exit-rate would never reach conclusive results within the desired A/B testing timeframe (as mentioned in Section 4.3). From a sales perspective, this might not seem that important if the design change yields a considerable increase in revenue. However, from a scientific perspective there needs to be compelling evidence that the change is not caused randomly.

One insight gained from this project was that focusing on high exit rate pages was not the ideal approach. However, it did give an idea of where the website might lack in quality or indicate places where there was room for improvement. We saw a few spelling mistakes, non-intuitive design, and unclear information present on some product pages. On Nordic Design Collective, artists maintained their own material and post their work themselves. They were allowed to post to the site without prior assessment, therefore there is a possibility of these issues reoccurring – even if they were removed.

The toughest and yet most important aspect of this project was to create a CRO procedure that would be effective. Moreover, these procedures will vary depending on the business, their conversion goals, and the user behavior within the website.

We learned that there are different approaches to evaluating the statistical significance of A/B tests. The choice of method depends on what data and subject area one is to collect and examine. The outcome of using an inappropriate test will result in improper conclusions. Knowing what variables are known will assist the user in determining which approach one could and could not use.

A suggestion for businesses that have recently begun their CRO journey and A/B testing is to take guidance from the decision tree in Figure 6-1, while focusing on the company's business objectives.

When to stop or continue an A/B test?

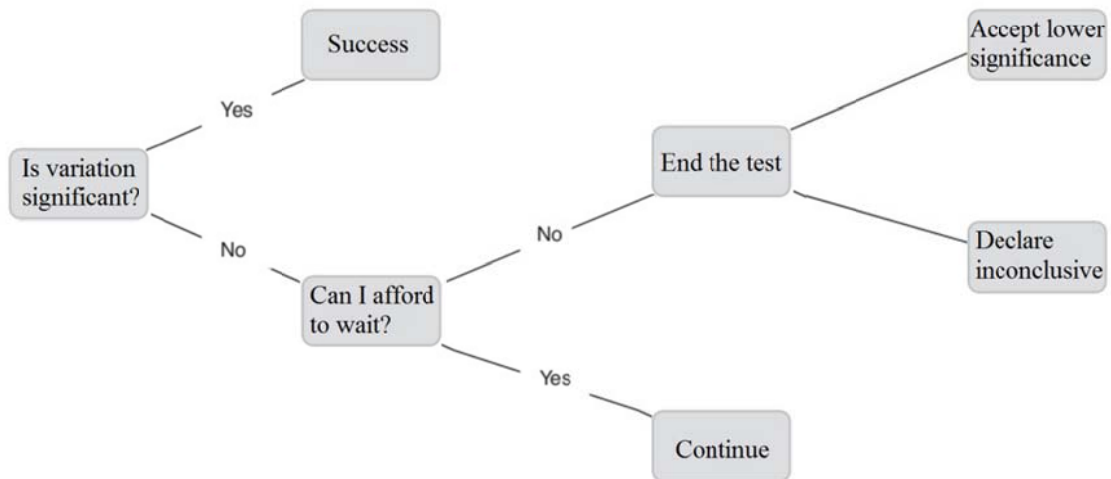


Figure 6-1: Decision Tree for A/B test culmination

It is tricky to know when to stop or continue an A/B test, and differs from business to business. If the test produces significantly positive results, then congratulations, make use of the difference intervals to see the benefits you can see. If the test has not reached significance, inspect the test's number of visitors remaining to reach significance value to evaluate if waiting is something you can afford. If yes, then let the test continue to run. If no, estimate the risk you take by using the difference intervals.

If the organization you are working with can:

- Iterate new design-proposals rather quickly,
- Run many experiments, and
- Has little downside risk of implementing inconclusive (yet improving) variations

then your organization is more likely to tolerate a higher error rate and accept a lower significance on tests.

Another suggestion prior to creating design proposals is to use the A/B testing tool's sample size calculator in order to roughly estimate the time needed to reach significance. Although this, as mentioned in Section **Error! Reference source not found.**, is not intuitive since it requires assumptions of a minimal detectable effect and knowledge of the baseline conversion rate –it can still save the experimenter from not wasting time and effort into a project that will not reach its

deadline. This is important in cases where projects that do not finish in time, are disregarded completely.

Some additional research that could have been done was usability testing and conducting surveys of users to get data directly from end users. This might end up producing different results than what we have seen from the data collected by Google Analytics and third-party tools. However, such surveys and similar data would have to be treated as a sampling of the whole population. This would also require finding a diverse group of individuals to replicate the demographics of the whole population. This would be both time-consuming and difficult unless there is some incentive to do this other research.

If we were allowed to do this project again we would focus on using more data to base our designs on. We would also like to implement measures to track users that have JavaScript disabled, cookie blocking, and other ad-blocking software. Although the market share of users having JavaScript disabled is very small, it should not be excluded. Additionally, being one of the few e-commerce sites that actually makes an effort to make a website usable with a pleasant user experience could bring a lot of attention amongst a niche group of users and could become very profitable in the future. There is currently support for tracking users without JavaScript enabled within Google Analytics through server-side libraries that are able to track all visitors; however this is not enabled by default. For an example of implementing this tracking see [42], as the Google Analytics documentation is quite confusing.

We would also have liked to go deeper into members' statistics and returning visitors to see if it was possible to find any patterns, flaws, or other issues that could be addressed to convert these users more efficiently. Perhaps an idea would be to find alternatives for encouraging the notion of membership with simple and easy sign-ups and sign-ins.

Furthermore, we would have wanted to use user engagement within the design process of the website. Although the HCI principles offer a good guide, in the process of interaction design designers collaborate with their target group(s) when creating a design. As the site already is very dependent on what the artists post on the website and these artists are in complete control of what they upload, it should not be a big step to involve them. As was discussed in Section 3.4.1, a small sample of people might not be an accurate representation of the website's population, especially when it comes to design, which is very subjective.

6.2 Limitations

One important limiting factor for us was time. As mentioned in Figure 6-1, if a test has not yet reached significance, one should evaluate the "visitors remaining" and see if one can afford to wait. According to Optimizely, if our detectable effect remained the same, we would only need to wait until 8400 more visitors interacted with the first A/B test (Figure 4-11) and around 300 more visits on the second test

(Figure 4-15). However, although we used this indication to wait and not stop our test, as said in Section 5.4.2, it did not help. However, it is reasonable to assume that our test would reach significance if we could afford to wait much longer.

Optimizely's reports were not as specific as one would have hoped when it comes to verifying their results by hand. Even though we used Optimizely's data when manually calculating confidence intervals, we still see a notable difference in results between the two approaches. To determine what causes the fluctuation in results, Optimizely's report should include: amount of false discoveries, the effect of continuous monitoring, and the significance change it caused.

Another limitation we faced, as mentioned in Section 5.4.1 (which we were also warned about in Table 3-1) concerned Nordic Design Collective not having full control of the webpage's source code. This made it difficult for us to fully understand some of the code and also made it difficult to implement critical code that only their third-party service provider had access to.

Nordic Design Collective is an e-commerce business meant to be used as a platform for Nordic artists to get their work to the consumers. The main goal of such a business is usually to generate maximum revenue at all costs. However, this is not their business model. Our design proposals had to take into account the integrity of their beliefs and that the web design should not deviate too much from that. We learned that it was a very fine line between the current design and what we were allowed to implement. Working with CRO and web design is very different from the norm in which their existing web designers work.

6.3 Future work

One thing we have not done that is important to know, especially for business owners, is to compare revenue accumulated from variations A and B. They are interested to know from the tests, in addition to the possible statistically accurate improvement, if they have an increase in revenue and thus can better validate the results of any changes. We did not manage to accurately measure this with the current setup and data available to us. The tracked click-through rate in our tests does not always equate to a micro/macro conversions owing to many potential factors, hence would be wise to validate using revenue accumulation for those pages that were tested.

One thing that should be done, if the website allows for it, is to fully integrate Optimizely with Google Analytics, as this is necessary and advantageous in order to achieve a more immersive and complete understanding of one's data and tests.

6.4 Required reflections

Web analytical tools give huge amounts of user- and behavioral information which can give a lot of insights concerning one's e-commerce site and where to start improving. Since there are all sorts of metrics, dimensions and filtering options

available on these tools, overwhelming the regular user, we found that the chosen metrics, dimensions and filters in this report are most useful and can be used concretely in a design-making process. Our conversion optimization approach is cost-efficient. The only expense is Optimizely, and the price of using its service depends upon the number of visitors one receives to the website. Startups tend to pay less than enterprises. There are other alternatives that will vary in price, features available, and in the technique used to collect and report data.

References

- [1] PostNord i samarbete med Svensk Digital Handel och HUI Research, 'E-barometern årsrapport 2015', 2015. [Online]. Available: http://www.hui.se/MediaBinaryLoader.axd?MediaArchive_FileID=48e5f483-55fo-4de2-aaeb-530bca0e1bao . [Accessed: 11-Nov-2016]
- [2] Paco Underhill, *Why we buy: the science of shopping*, Updated and rev. New York: Simon & Schuster Pbks, 2008, ISBN: 978-1-4165-9524-3.
- [3] Dan Saffer, *Designing for interaction: creating innovative applications and devices*, 2nd ed. Berkeley, CA: New Riders, 2010, Voices that matter, ISBN: 978-0-321-64339-1.
- [4] Bryan Eisenberg and John Quarto-von Tivadar, *Always be testing: the complete guide to Google website optimizer*. Indianapolis, Ind: Wiley Publishing, 2008, Serious skills, ISBN: 978-0-470-29063-7.
- [5] Dan Siroker and Pete Koomen, *A / B Testing: The Most Powerful Way to Turn Clicks Into Customers*. Wiley, 2015, ISBN: 978-1-118-79241-4.
- [6] John Ekman, *Konvertering – Vad är det?* [Online]. Available: <http://www.conversionista.se/konvertering/> [Accessed: 07-Nov-2016].
- [7] Khalid Saleh and Ayat Shukairy, *Conversion optimization*, 1. ed. Sebastopol, Calif: O'Reilly, 2010, ISBN: 978-1-4493-7756-4.
- [8] Monetate, Inc., 'The Ecommerce Quarterly Benchmark Report for Q3 2016', 2016. [Online]. Available: http://info.monetate.com/EQ3_2016-benchmarks.html . [Accessed: 11-Nov-2016]
- [9] Optimizely, 'Take action based on the results of an experiment - Optimizely Knowledge Base', 13-Sep-2016. [Online]. Available: https://help.optimizely.com/Analyze_Results/Take_action_based_on_the_results_of_an_experiment . [Accessed: 11-Nov-2016]
- [10] A. Marcus, Design, User Experience, and Usability: User Experience Design for Everyday Life Applications and Services: Third International Conference, DUXU 2014, Held as Part of HCI International 2014, Heraklion, Crete, Greece, June 22-27, 2014, Proceedings. Springer, 2014. pp. 359
- [11] A. Cooper, R. Reimann, D. Cronin and A. Cooper, *About face 3*, 1st ed. Indianapolis, IN: Wiley Pub., 2007.
- [12] Urban Lindstedt and Lisa Bjerre, *E-handlarens handbok: vägen till framgångsrik e-handel*. Stockholm: No Digit Media, 2009, ISBN: 978-91-976062-3-3.
- [13] Douglas K. Van Duyne, James A. Landay, and Jason I. Hong, *The design of sites: patterns, principles, and processes for crafting a customer-centered Web experience*. Boston: Addison-Wesley, 2003, ISBN: 978-0-201-72149-2.
- [14] Christian Holst, 'Fundamental Guidelines Of E-Commerce Checkout Design', *Smashing Magazine*, 06-Apr-2011. [Online]. Available: <https://www.smashingmagazine.com/2011/04/fundamental-guidelines-of-e-commerce-checkout-design/> . [Accessed: 11-Nov-2016]
- [15] Dibs Payment, 'Svensk E-handel - Dibs Årliga Rapport Om E-handel 2015', 2015. [Online]. Available: http://www.dibspayment.com/sites/corp/files/files/SE/NEH_SE_2015_web.pdf?ga=1.115718211.479219750.1475240058 . [Accessed: 11-Nov-2016]
- [16] "About data sampling - Analytics help," [Online]. Available: <https://support.google.com/analytics/answer/2637192?hl=en> . [Accessed: 01-May-2017]
- [17] "Google Analytics Usage Statistics." [Online]. Available: [//trends.builtwith.com/analytics/Google-Analytics](http://trends.builtwith.com/analytics/Google-Analytics) . [Accessed: 28-Dec-2016].

- [18] "Enhanced Ecommerce reports", Google, [Online]. Available: <https://support.google.com/analytics/answer/6014872?hl=en> . [Accessed: 15- Nov-2016].
- [19] J. Phillips, *Ecommerce Analytics: Analyze and Improve the Impact of Your Digital Strategy*, 1st ed. New Jersey: Pearson Education, Inc., 2016.
- [20] "How a web session is defined in Analytics – Analytics Help." [Online]. Available: <https://support.google.com/analytics/answer/2731565?hl=en> . [Accessed: 01- May-2017].
- [21] "Optimizely: Optimize digital experiences for your customers." [Online]. Available: <https://www.optimizely.com/ab-testing/> . [Accessed: 16-Dec-2016].
- [22] E. Guba, *The Paradigm dialog*, 95th ed. Newbury Park, Calif.: Sage Publications, 1990.
- [23] Datanyze, "Testing and Optimization Market Share Report | Competitor Analysis in Alexa top 10K | Optimizely, Visual Website Optimizer, OptinMonster," Datanyze. [Online]. Available: <https://www.datanyze.com/market-share/testing-and-optimization/> . [Accessed: 14-Nov-2016].
- [24] Optimizely, "Introducing Optimizely X: The Experimentation Platform for Marketers, Developers and Product Managers to Experiment Everywhere." [Online]. Available: <http://www.prnewswire.com/news-releases/introducing-optimizely-x-the-experimentation-platform-for-marketers-developers-and-product-managers-to-experiment-everywhere-300328587.html> . [Accessed: 14-Nov-2016].
- [25] "When To Do Multivariate Tests Instead of A/B/n Tests," ConversionXL, 07–07-Sep-2015. [Online]. Available: <http://conversionxl.com/multivariate-tests/> . [Accessed: 28-Dec-2016].
- [26] "When to Run Bandit Tests Instead of A/B/n Tests," ConversionXL, 14–14-Sep-2015. [Online]. Available: <http://conversionxl.com/bandit-tests/> . [Accessed: 28-Dec-2016].
- [27] Tom Tullis and Bill Albert, *Measuring the user experience: collecting, analyzing, and presenting usability metrics*. Amsterdam ; Boston: Elsevier/Morgan Kaufmann, 2008, The Morgan Kaufmann interactive technologies series, ISBN: 978-0-12-373558-4.
- [28] Alex Birkett, 'A/B Testing Statistics Crash Course: Ignorant No More', ConversionXL, 01-Oct-2015. [Online]. Available: <http://conversionxl.com/ab-testing-statistics/> . [Accessed: 11-Nov-2016]
- [29] Jakob Nielsen, 'Putting A/B Testing in Its Place', NN/g Nielsen Normal Group, 15-Aug-2005. [Online]. Available: <https://www.nngroup.com/articles/putting-ab-testing-in-its-place/> . [Accessed: 11-Nov-2016]
- [30] "Implement the snippet for Optimizely Classic," Optimizely Knowledge Base, 14-Apr-2016. [Online]. Available: https://help.optimizely.com/Set_Up_Optimizely/Implement_the_Optimizely_snippet . [Accessed: 01-May-2017].
- [31] "Adding analytics.js to Your Site | Analytics for Web (analytics.js)," Google Developers. [Online]. Available: <https://developers.google.com/analytics/devguides/collection/analyticsjs/> . [Accessed: 28-Dec-2016].
- [32] S. Tonkin, C. Whitmore, and J. Cutroni, *Performance Marketing with Google Analytics: Strategies and Techniques for Maximizing Online ROI*. John Wiley and Sons, 2011.
- [33] L. Pekelis *et al.*, "The New Stats Engine," *Optimizely, Department of Statistics, Stanford University*, 20-Jan-2015. [Online]. Available: http://pages.optimizely.com/rs/optimizely/images/stats_engine_technical_paper.pdf . [Accessed: 14-Nov-2016]

- [34] Optimizely, “Optimizely Stats Engine: An overview and practical tips for running e...,” 18:47:21 UTC. [Online]. Available: <https://www.slideshare.net/optimizely/optimizely-stats-engine-webinar> . [Accessed: 01-May-2017]
- [35] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)*, pages 289–300, 1995.
- [37] M. Abraham, A. Lipsman and C. Meierhoefer, "The impact of cookie deletion on the accuracy of site-server and ad-server metrics: An empirical comScore study," *comScore*, Jun. 6, 2007. [Online]. Available: <http://www.comscore.com/Insights/Presentations-and-Whitepapers/2007/Cookie-Deletion-Whitepaper> . [Accessed: 01-May-2017].
- [38] comScore, "The Impact of Cookie Deletion on Site-Server and Ad-Server Metrics in Australia," *comScore*, 2011. [Online]. Available: <http://www.comscore.com/Insights/Presentations-and-Whitepapers/2011/The-Impact-of-Cookie-Deletion-on-Site-Server-and-Ad-Server-Metrics-in-Australia-An-Empirical-comScore-Study> . [Accessed: 01-May-2017].
- [39] "The cost of ad blocking—PageFair and Adobe 2015 Ad Blocking Report ", Aug 10. 2015. [Online]. Available: http://downloads.pagefair.com/reports/2015_report-the_cost_of_ad_blocking.pdf . [Accessed: 01-May-2017].
- [40] GfK, "Understanding cross-device usage to optimize digital campaign planning—Multi—device case study", "GfK, 2015. [Online]. Available: http://www.gfk.com/fileadmin/user_upload/website_content/Succes_Stories_PDF/Global/Facebook_multi_device_case_study_long_version.pdf . [Accessed: 01-May-2017].
- [41] “Revenue tracking goals in Optimizely Classic,” *Optimizely Knowledge Base*, 14-Apr-2016. [Online]. Available: https://help.optimizely.com/Measure_success%3A_Track_visitor_behaviors/Revenue_tracking_goals . [Accessed: 01-May-2017].
- [42] B. Jovanović, “How to Track Website Visitors with JavaScript Disabled,” Moz. [Online]. Jun. 2013. Available: <https://moz.com/ugc/how-to-track-website-visitors-with-javascript-disabled> . [Accessed: 01-Dec-2016]
- [43] M. Thörn, “Konverteringsoptimering för e-handelsbutiker – En möjlighet för e-handelsföretag att identifiera sina brister och styrkor”, Master’s thesis, Dept. Comput. Sci. and Commun., KTH., Stockholm, Sweden, 2012 [Online]. Available: http://www.nada.kth.se/utbildning/grukth/exjobb/rapportlister/2012/rapporter12/thorn_marcus_12065.pdf . [Accessed: 04-April-2016]
- [44] M. Lundvall, “Konvertering och användbarhet – En undersökning om hur besökare på webbplatser med e-handel kan konverteras till kunder”, Master’s thesis, Dept. Comput. Sci. and Commun., KTH., Stockholm, Sweden, 2011 [Online]. Available: http://www.nada.kth.se/utbildning/grukth/exjobb/rapportlister/2011/rapporter11/lundvall_martin_11061.pdf . [Accessed: 04-April-2016]

Appendix A: Statistical Significance Calculation Using the Null Hypothesis.

The Null Hypothesis

H_0 : $\Delta = 0$, that there is no difference between the conversion rates of both designs. Will it however show that the probability of our data is less than 5% probable to obtain, then we will reject H_0 and decide for H_1 .

H_1 : $\Delta \neq 0$, that there is a systematical difference in conversion rates.

C_O = Conversion rate of the original design (conversions/visitors)

C_V = Conversion rate of the varying design (conversions/visitors)

Let C_O be a random variable that describes the conversion rate of the original design and let C_V describe the conversion rate of the experimental design. Let $D = C_V - C_O$ be a random variable that describes the difference between the conversion rate of the variation and the original. Then, we have that

$$\bar{D} = \frac{1}{N} \sum_{i=1}^N C_V_i - C_O_i$$

Will describe the mean of the difference in conversion rates of our sample data, where N is the data distributed into total days of the test running.

The population variance is then defined as:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^n (D_i - \bar{D})^2,$$

And the population standard deviation being $\sqrt{\sigma^2} = \sigma$, which will then be used to calculate our sample standard deviation by:

$$\sigma_{\bar{D}} = \frac{\sigma}{\sqrt{N}} \quad (1.1)$$

All variables needed to obtain a z-score is now acquired. We have our observed sample mean and our sample standard deviation, along with a null hypothetical assumed conversion rate difference of $\mu_{\bar{D}} = 0$.

To determine significance of our results and to see how many standard deviations away from the assumed conversion rate mean we are, we make use of z-scores, which is defined as:

$$Z = \frac{\bar{D} - \mu_{\bar{D}}}{\sigma_{\bar{D}}} \quad (1.2)$$

We are going to use a two-tailed test since we are seeking for probability limits at both ends of the spectrum (either significantly positive or negative).

If the z-score is higher than 1.96, it is then less than 5% probable to obtain the observed data, meaning, we reject H_0 and conclude H_1 , that there exists a systematic difference in in the conversion rates. The experimental design has brought a **positive** impact.

If the z-score is less than -1.96, it likewise less than 5% probable to obtain the observed data, we reject H_0 and conclude H_1 . The experimental design has brought a **negative** impact.

If the z-score lies between -1.96 and 1.96, no decision based on statistical significance can be made. The difference in conversion rates can purely be due to chance.

For a 90% significance in a two-tailed test, the z-score threshold lies at ± 1.6449 . A score above would give a positive impact and a score below would give a negative impact.

The confidence interval is calculated by:

$$\bar{D} \pm 1.96 \times \sigma_{\bar{D}} \quad (1.3)$$

Appendix B: Detailed results

A/B Test 1 Results: Conversion rate difference over the days running.

-0.0118, 0.0282, 0.0298, 0.03, -0.013, 0.0271, 0.003, -0.0103, 0.0472, 0.0017, 0.0139, 0.0047, 0.0105
 -0.0014, 0.0091, 0.007, 0.0279, -0.0166, -0.0256, 0.0331, -0.0092, -0.0166, 0.0159, 0.0205, 0.0178,
 -0.0112, 0.0193, -0.0094, 0.0163, -0.0159, -0.0035, -0.0178, -0.0142, -0.0159, -0.0378, 0.0069,
 0.0372, -0.0258, -0.023, 0.0196, -0.0182, 0.0211, -0.0356, 0.023, -0.0398, 0.0031, 0.014, 0.0038,
 0.0147, 0.0045, -0.0007, -0.0045, 0.0199, 0.0149, 0.0152, -0.005, 0.0068, -0.003, 0.0153, 0.0046,
 0.0034, 0.0009, -0.007, -0.007, -0.0117, -0.0068, 0.0437, 0.0078, 0.0231, 0.1065,

Mean of differences, $d = 0.005067$

$N = 70$

Variance = 0.00048685

Standard deviation = 0.022065

Sample standard deviation = 0.002637

$Z = 1.9215$

A/B Test 2 Results: Conversion rate difference over the days running.

0.1071, 0.0904, -0.0939, 0.1161, 0.0545, 0.1875, 0.1, 0.1429, 0.2201, -0.1649, -0.0229, 0.1339,
 0.2361, 0.0718, -0.0636, 0.4524, -0.2250, 0.0461, 0.2, 0.3129, 0.00429, 0.1273, -0.2, -0.1889, 0.0167,
 0.4514, -0.1786, 0, -0.111, 0.4332, 0.1643, -0.0364, -0.0364, -0.2292, 0.0192, -0.1333, 0.3455, -0.191,
 -0.1495, 0.0324, -0.0972, -0.0660, 0.556, -0.0795, 0.4222, 0.0286, -0.0392, 0.0875, 0.175, 0.0304,
 0.0682, -0.1804, 0.2243, 0.2037, 0.3125, 0.1438, -0.2468, -0.2468, -0.2232, 0.2108, 0.1591, 0.1622

Mean of differences, $d = 0.0596$

$N = 60$

Variance = 0.0323

Standard deviation = 0.1797

Sample standard deviation = 0.0232

$Z = 2.569$

TRITA-ICT-EX-2017:36