CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

Towards extending A/B Testing in E-Commerce sales processes

A Master's thesis in Software Engineering

KONSTANTINOS KOUKOUVIS

ROBERTO ALCAÑIZ CUBERO

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2015

Master's thesis 2015

Towards extending A/B Testing in E-Commerce sales processes

KONSTANTINOS KOUKOUVIS
ROBERTO ALCAÑIZ CUBERO

Department of Computer Science and Engineering
Division of Software Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2015

Towards extending A/B Testing in E-Commerce sales processes
KONSTANTINOS KOUKOUVIS & ROBERTO ALCAÑIZ CUBERO

Supervisor: Patrizio Pelliccione, Department of Computer Science and Engineering
Examiner: Richard Berntsson Svensson, Department of Computer Science and Engineering

Master's Thesis 2015
Chalmers University of Technology
University of Gothenburg
Department of Computer Science and Engineering
Division of Software Engineering
SE-412 96 Gothenburg
Sweden
Telephone +46 31 772 1000

Typeset in LaTeX
Gothenburg, Sweden 2015

Towards extending A/B Testing in E-Commerce sales processes
KONSTANTINOS KOUKOUVIS
ROBERTO ALCAÑIZ CUBERO
Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg

Abstract

E-commerce has traditionally been a field where online controlled experiments, such as A/B testing, take place. While most of these tests focus on evaluating different visuals (graphics, colors, layouts) this thesis is about using them to evaluate different sales processes. The authors began by developing a tool for a local company designed to empower businesses in helping their customers understand their needs and make decisions on purchasing their products or services. Then an embedded case study took place, which involved the evaluation of the tool by company personnel and also experiments under real life conditions. The goal of this thesis is to prove whether A/B testing can be applied in E-commerce sales processes and the formation of a set of guidelines designed to help business with interest of conducting similar experiments.

Keywords: A/B Testing, web testing, decision assistant, sales process.

Acknowledgements

A big thanks to our supervisor for all his help and our families and friends for their support throughout the carrying out of this Master Thesis.

Konstantinos Koukouvis & Roberto Alcañiz Cubero, Gothenburg, 2015

Contents

1
Introduction

Around the early 2000s Greg Linden at Amazon created a prototype to show personalized recommendations based on items in a user's shopping cart [1]. When a user would add an item a new recommendation would show up. Each item added would garner different recommendations. While the prototype looked promising according to Linden, the marketing senior vice president was dead set against it claiming that it would distract the users from actually buying the items they would buy. Regardless of this, Linden went on to run a controlled experiment; the feature proved to be an immediate success leading to the launch of shopping cart recommendations, which since then has been used by multiple similar sites.
The world-wide-web has evolved into the perfect playground for evaluating different ideas through controlled experiments. This type of experiments can be called by different names with only slight differences between them, such as randomized experiments, A/B tests, split tests, Control/Treatment tests, MultiVariable Tests (MVT) and parallel flights. One of the earliest examples of running an online A/B test was the aforementioned experiment carried out by Linden [1]. The controlled experiments embody the best scientific design for establishing a causal relationship between changes and their influence on user-observable behavior [2]. In the context of online experimentation, A/B testing can be utilized in order to analyse user behaviour. The idea behind A/B testing in an online environment is to create two variants of a single website and then randomly assigning to each visiting user one of the variants. Those variants are usually called:

1. The Control, which is usually the currently existing version and
2. The Treatment, which is usually a new version that needs to be evaluated.

The variants could also be two completely new versions of a service. Different metrics can be set up for those tests with interests that can range from runtime performance to implicit and explicit user behaviors. The results of those tests, as well as possibly some survey data (by prompting visiting users to fill questionaires evaluating their interaction with the website), are collected. Afterwards statistical tests are conducted in order to evaluate the existence of statistically significant differences between the two variants, thus leading to the acceptance or rejection of the null hypothesis, which is that there is no difference between the two versions [3]. A key issue is that the users should have a consistent experience for the service, that is they should always see the same variant when coming back to the service.
As Kohavi et al. say: "Most organizations have many ideas, but the return-on-investment (ROI) for many may be unclear and the evaluation itself may be expensive. [...] A live experiment goes a long way in providing guidance as to the value of

the idea." [2]. This experimentation however is mostly based in altering the usability aspects of a website. Also citing Bosch "In practice, however, experimentation in online software is often limited to optimizing narrow aspects of the front-end of the website through A/B testing [...]" [4]. In this master thesis the conducted A/B tests were solely focused on the sales process and not on a purely visual aspect.

In the "traditional" sales process the customer would enter a store and then receive the help of the owner or that of a store worker. What follows then is a series of questions and answers that help the seller gain information about the customer's needs and help him realize the best way to market his product. As stated by Monty the sales process is like a dance, where the seller and the buyer must take some steps [5]. The sales process is defined by the Plan stage, where the seller must strategize and gain knowledge over his customer, the Qualify stage, where the seller must decide whether the customer is worth his time and guide the conversation slowly toward his solution based on the customer's needs. Then comes the Prove stage, where the seller is providing a demo of the product, leading to the Close stage where the seller is taking an actual order for his product. After a successful sale come the Implement and Measure stage where the seller is ensuring his product runs smoothly for the customer and measures the results after implementation.

In the virtual setting of E-commerce though several of these stages are not possible. The customer browses the website on his own, checks the information that is given in the product's page and then has to figure out a way to decide if the specific product suits his needs. The Decision Assistant that the authors have developed for the company Sonician during the time of this master thesis has been designed to do exactly that. Provide the business owners with an interface where they would develop their own virtual assistant for their customers in order to gain knowledge on their needs and provide them with the product that fits them. In order to do this they can use the Decision Assistant to create a process which can consist of a series of steps which their customer has to go through. These steps are highly customizable, thus bearing the potential of creating several processes with different questions, number of steps and/or required information from the customer. This complex system offers an ideal base to conduct testing over the process itself, instead of conducting traditional user interface or usability tests, for instance in terms of testing simpler processes against more complex ones.

Once the tool was developed, the point to address to check if the authors could create some kind of experimentation in order to be able to achieve best practices for creating sales processes. The ultimate base for this experimentation would be the use of this tool. Given the nature of this tool and of A/B testing, the authors decided to study the suitability of implementing this kind of experimentation in order to construct and evaluate best practices in sales processes

Summing up, this thesis is going to address the issue of whether A/B testing can be used to evaluate and improve the success of a sales process in an online decision support system. In order to conduct this research, a case study was developed, which consists of a round of interviews and then a set of controlled experiments were carried out. The first part of the study provided a set of metrics to customize the decision assistant tool, which was used to host the desired experiments (A/B tests). The customizations were set as variants of a sales process, and the results

of these tests indicated whether there is a significant difference in the use of one variant over the other.

1.1 The Company: Sonician

Before proceeding further, a few more words about the company where the authors did their Master Thesis at are required. Sonician, established in 2008, consists of a team of dedicated and passionate people from different backgrounds, all with their own special abilities and experiences, brought together by a common vision and strong level of engagement and enthusiasm. The company's focus is on Marketing Automation where they have launched and maintain their own software system, Otto. With the Decision Assistant tool Sonician's vision was to improve the user experience of a lot of commercial sites by creating a tool that can help the site's visitors identify and satisfy their needs in the most suitable way possible. Sonician, henceforth referred to as 'the company', assigned the creation of that tool to the authors of this Thesis.

1.2 Thesis Outline

The remainder of this thesis is structured as follows: In the next Section, the background and related work are presented. In Section 3, there is a detailed description of the research approach followed. Section 4 is dedicated to the description of the tool developed, as well as the case study conducted, including its setup and results. Section 5 reports about the experiments conducted with the developed tool, and presents the results of these. The thesis ends with the conclusions chapter, Section 6, which includes discussion and future work.

2

Background And Related Work

Since this thesis regards performing a controlled online experiment, the current status of experimentation on the web is going to be reported while highlighting the status of A/B testing, which is the specific method used in this thesis. Also since this type of controlled experiment is going to be used on testing its suitability to conduct a Sales process in the context of a decision assistance framework, information on the evolution and status of the sales process and decision support systems is going to be conveyed.
The following chapter provides a summary of the background work and a review of the current work related to the previously stated areas and it is structured as such:

1. The sales process (section 2.1),
2. Decision support systems (section 2.2),
3. Controlled experiments in the web and especially A/B testing (section 2.3).

In addition to provide the current state of the art on the aforementioned concepts, and since the authors believe that those topics might require a strong technical understanding of related concepts, clarifications on the terminology used are also included in this section. All explanations can be traced from the same papers utilized in the state of the art review.

2.1 The Evolution Of The Sales Process

The sales process has been continuously evolving. As customers become more sophisticated and better-informed the sales process is much less about selling a product and much more about creating a relationship [6]. And as Vargo and Lusch point out there has been an accompanying shift from product to service selling (otherwise known as servitization) in many business-to-business interactions [7]. Thus, according to Storbacka et al. the constituency that is involved in the sales processes has been inevitably widened [8].
The advancement of the web has also contributed to the changes in the sales process. First and foremost since more and more of the stores create an online presence their sales tactics must be adapted to fit the new medium. This gave birth to E-commerce which was defined by Chen and Jiang as “transactions conducted by buyers who perform online interactions with sellers and their final decisions are influenced accordingly, no matter which type of channel they choose to transact

with at last" [9]. The evolution of web 2.0 also has brought a wave of social websites to the forefront of the web. A deep investigation on how each different stage of the sales process can be influenced by the social media is has been carried out by Andzulis (et al.) [10].
With the establishment of an online presence the companies turned to adding significant technology in the sales process. The most important of which was it's automation. Erffmeyer and Johnson published a study where they explored the company practices [11]. The result of that study showed that in most cases the organization's goals behind the automation of the sales process were the need to improve the organization's efficiency.
This evolution sparked the coining of the term E-Selling by Parvinen et al. which is defined as "a human-like interaction directed at increasing customer value by securing a business exchange" [12]. In their paper they continue on to compare the real-life selling process with its equivalent virtual one. One aspect for example is that of tailoring proposals on a verbal level. In the virtual context this can be seen in the use of recommender applications (more common in B2C) and configurators (more common in B2B). Amazon.com and construction companies have served as best practices [13] [14] [15]. There are also some online games that include some 3D shop's recreations which tailor their in-store approaches. One of the most advanced approaches is to use click stream data to infer customers' cognitive types and tailor the website's look-and-feel and marketing content accordingly [16]. They point out however that interactive e-selling applications with focus on securing a deal are very scarce. Another identified issue is that in the context of the online store there is nobody around to interact with the customer through the purchase process especially when the urge to hesitate occurs, thus leading the customer to leave the store without making a purchase.
Another aspect that Parvinen et al. also point out is trust [12]. Economic theory posits the value of sales interaction on the beneficial information exchange and under imperfect information, the trust between seller and the buyers [17]. In the E-selling context the research on trust can be divided into three strands [18] [19] [20]:

1. Trust in digitally mediated human-to-human interaction,
2. Trust in human-to-computer interaction and
3. Comparisons of human-to-human and human-to-computer trust.

In the area of virtual sales assistants and recommender applications it can be seen that some of the processes of building trust have successfully been utilized. The aforementioned recommender applications can be different in design depending on whether creating trust or alleviating distrust is key [20]. Kakavelakis also pointed out that the extent to which humans can be trained into instrumental empathy is significantly limited [21].

2.2 The Status Of Decision Support Systems

Although customer decision support systems have long been viewed as a promising selling tool, until recently, they have had very limited application [22]. Now, in the

web 2.0 era of self-service, a system of that caliber can offer a significant value to the entire sales process, guiding the customer in order to reach a buy-or-not decision [23]. Those systems find heavy use in the context of product configuration. A study on that context showed that customer satisfaction can be determined in part by the ease or difficulty of the process the customer follows [24]. Another study showed that the customer was willing to pay more for a product when the effort to evaluate it took less, especially in cases where the customer is less skilled [25]. In any case the use of such systems can facilitate a solution driven approach to marketing [22]. For any good or service one of the most important factors that can guide customers to solutions most appropriate for their situations is an assessment of needs [22]. By identifying those needs product options can be better determined and the solution can be more appropriately configured. In addition, even if the customers understand their needs, they might not be able to successfully judge the best product choice. Thus, as Grenci and Todd conclude "the Web can offer more than just promotion, electronic order forms, and data transfer... it can facilitate the use of decision support tools that actually configure or define the product based upon individual customer needs, thus guiding the purchase process toward effective solutions" [22].
Yang et al. proposed a novel decision support framework that organized natural language questions into an analytic hierarchy in order to produce a final decision. It was tested in a target validation environment for the medical industry, but also in a product recommendation environment using review texts as input. The results of their experiments showed generally lower values in the product recommendation than in the target validation evaluation which means that when it comes to the everyday decision problem that might seem easy, it is actually harder than a scientific decision based on technical publications [26].
In another study, Yu proposed a web-based consumer-oriented intelligent decision support system for personalized E-services in the field of E-tourism and E-investment. Through their system all the customer decision-making process steps can be supported. According to him, comparing either the E-tourism or the E-investment prototype of his system to any current domain-specific web site, it seems that "the prototypes are the only systems that are capable of supporting consumers fully in selecting, designing, and implementing personalized decision plans." [27].
Lastly Qaed and Sutcliffe suggested an adaptive decision support system for E-commerce which matches the appropriate tool support and decision strategy advice to the user's preferences and motivations. They concluded that "the use of a social actor in delivering system advice is effective and persuasive. In addition different decision tools are needed for users to allow them to perform their decision strategies in different scenarios and product domains. These different decision tools and advice could support users in making their decisions more effectively." [28].

2.3 Controlled Experiments In The Web

Davenport expresses the importance of testing as a tool to make tactical decisions in a range of business settings, from banks to retailers to dot-coms, and stresses the need to create a testing mind-set in companies in order to move testing "out of the laboratory and into the boardroom" [29].

In the online environment the goal of hosting controlled experiments is to perform the evaluation of new ideas in order to try and find out if those new ideas will grant any benefits compared to the previous arrangement of the system when applied [30]. As this is usually performed by companies or organizations, the ultimate benefit of conducting experiments is to increase return-on-investment [3]. This matter has been widely discussed in scientific literature, with the particularity that for web development this technique has been traditionally used to evaluate user experience aspects.

2.3.1 A/B Testing

There has been some initial research into how analytics, including A/B testing, could be leveraged in software product planning. The analytics allow grounding the decisions on evidence, hence they limit the influence of power and politics. A paper by Fotrousi et al. suggests shifting from monitoring web pages to monitoring features in order for analytics to become effective for product planning [31].
A/B Testing has been widely used in web development as a kind of controlled experiment. In order to perform these experiments, users are randomly presented with either the original version of the website (referred to as the Control), or a variant that represents the new idea to test (referred to as the Treatment). When it comes to A/B testing in a Web environment, this has been applied to aspects concerning user experience or interaction with the website [3] [32] [33] [30] [34] [35].
Also A/B testing has been discussed in a lot of recent business literature such as Sauro and Lewis [36], Finger and Dutta [37], and Seufert [38], while Clark indicates that the methodology has been adopted in the mobile gaming industry, something which proves that A/B testing can be feasible in more than just the web domain. In order for the analysis data to be collected though there is a need for an online presence of the client at some point in time [39].
One of the important ideas of A/B testing is that for it to be properly conducted, it needs to be a random experiment. This means that the participants should be randomly assigned to either the Control or the Treatment [3]. In order to achieve this, two techniques are used, namely Randomization Algorithms and the Assignment Method.
The Randomization Algorithms will map users to one of the variants. It is of great importance to use a good algorithm since as stated before, the experiments must be random, so the chosen algorithm must guarantee that there is no bias towards any variant. Another important aspect of the algorithm is that it should map single users to the existing variants instead of mapping individual sessions (that is, every single access to the site regardless of if it is performed by a totally new user or by one who has visited the site before) to variants. This means that a user who enters the site several times should be redirected to the same variant. Finally, the algorithm must support Ramp-up. Ramp-up is defined as a characteristic in which the percentage of users assigned to the Treatment can be steadily increased, so that typically the experiments will start by sending perhaps a 0.1% of the users to the Treatment, and gradually increase this size to finally achieve a 50% balance.
The Assignment Method works as a load balancer (a technique which distributes

workloads to different resources) that enables the system to execute the correct path to the variant selected for the user with the Randomization Algorithm.
The idea of expanding A/B testing in online services from its initial (and somehow) traditional perspective of user experience has also been recently proposed by Hynninen and Kauppinen [34]. In this study they draw conclusions on how A/B testing appears to be a promising method in customer value evaluation. In this context, the key factor to success would be to properly set the customer metrics to evaluate. However, A/B testing has its limitations. Among those, it has been concluded that A/B testing requires a large number of users involved in the experimentation [3] [33]. This is especially true when dealing with variances that may be experienced by only a small share of the website users. Another big factor is related to technical issues. It might actually be difficult to implement the testing infrastructure, since the experiments might involve an unusually large collection of data which must be conducted in a reliable way. Choosing the correct metrics beforehand, is a critical factor when it comes to making right decisions over product planning. Finally, A/B testing as it is, only provides information on how users behave, but it lacks all information about why. It has been suggested that these limitations can be partially solved by combining A/B testing with some sort of qualitative user research methods, "especially in contexts where there is large variability between users" [34].

2.3.2 Analyzing A/B Test Results

There are also references in literature about the outcomes of A/B Testing, and how to analyze those results. More specifically, the related work in this aspect is linked on how to successfully conduct statistical analysis on the gathered data in order to draw conclusions [35] [40].
It is worth noting that A/B testing may produce unexpected results, and researchers must know how to cope with them. According to Kohavi, almost all unexpected results in A/B Testing are due to errors that are not easy to predict unless there is a specific intention from the researchers to provoke them [32]. One suggestion for online experiments is to use extensive A/A experiments.
A/A testing, also referred to as Null Test, shares many commonalities with A/B Testing, with the key difference being that the Control and the Treatment are exactly the same, thus all users will see the same variant of the website. The idea behind this kind of test is to test the experimentation system [32] [40].
By utilizing such A/A tests results and conducting data quality checks the experimenters can detect certain problems. It is pointed out however, that if a result seems to be strange or unexpected the most probable cause is that the experimenters failed to understand particular behaviours by the users, or perhaps design problems exist either in the experiment or in the software itself.

3
Research Approach

3.1 Research Purpose

The purpose of the study is to extend the traditional vision of A/B testing in online environments, particularly in the E-commerce environment, by trying to prove its suitability not only in testing for visual differences but also on evaluating the correct approach when it comes to choosing between sales processes, specifically in the business-to-business or business-to-consumer perspective of automated sales.

3.2 Research Questions

A/B testing is being used in more and more online systems these days to maximize their business value and their return of investment, but this kind of tests are usually performed on experiments altering usability related aspects of the website. Citing Bosch [4] "In practice, however, experimentation in online software is often limited to optimizing narrow aspects of the front-end of the website through A/B testing [...]". In light of evaluating this method on its effect on the sales process, the following research questions were defined:
RQ1: How can the use of A/B testing be extended from visual aspects of online services in order to optimize sales processes in the E-commerce domain?
RQ2: What could a generalized framework, produced from the aforementioned use of A/B testing, contain so as to be adapted by companies in the field?

3.3 Research Methodology

In order to answer the proposed research questions, and taking into account the complexity of the study, the authors proposal is to divide this study into two units of analysis. Those units, which are interrelated, are a case study followed up by a series of experiments. Both units deliver different types of data, but the outcomes of the case study are the foundational base of the experiments. The rationale on why to conduct this two units of analysis is explained in the section Data Colection. However, the scope of this thesis is not limited to the study proper, but it is enhanced by the development of a tool onto which the experimentation will be conducted. The conclusions of the study were drawn from the outcome of experiments that were conducted with real customers and associates of the company the authors developed the tool for. The first unit consists of a case study, while the second

consists of an experiment. Taking into account the development period, the whole structure of this Master Thesis comprises of the following four phases:

3.3.1 Phase One: Requirements And Development

During this phase the authors worked closely with the company in order to elicit the needed requirements for the implementation of the suggested tool. After that and for the next three months the implementation of the tool followed using the Laravel PHP framework. Along with the development, a literature review was also carried out by the authors, the results of which can be seen in the second chapter of this thesis. Sites such as Google Scholar and IEEE Xplore were used to create a list of sources, and then from them, by utilizing the snowballing method, expand the knowledge base.

3.3.2 Phase Two: A Case Study

The case study involved the use of the tool under real conditions. Upon a meeting and a brainstorming session with employees of the company, it was agreed that no less than 6 people would participate in this phase. The tool was given to employees of the company in order for them to produce their own website which contains their decision assistant for the sale of a hypothetical product or service. The role of the authors in this phase was merely observational, providing only certain training and help with technical functionalities of the tool. Each decision assistant created had in turn two variants. The differences in those variants were only in matter of the sales process, either technical or visual, but always with heavy impact on the process. Each person responsible for the decision assistants was enquired about the metrics/parameters that he/she found more valuable in order to conduct the follow up experiments. These metrics represented intended business goals, since the purpose was that those employees acted as if they were responsible for a particular company or business. The selected metrics are presented in 5.2.4. The results of this case study, showcased in the following chapter, were used to evaluate the behaviour of the potential users of the tool.

3.3.3 Phase Three: Experiments

The Decision Assistant flows that were created in the previous phase were then put to the test. Using the company's client base each decision assistant was sent out to a varying number of clients and contacts, ranging from 83 to 8272 persons per experiment, who then proceeded to run the process and follow it up until the point each of them chose. A random number of those users was in each case given either variant A or variant B. The variants used are presented in 6.1. Using the metrics and characteristics defined in the previous phase the results of each experiment were tested for any significant difference between their two variants. After a sufficient time period ranging from two weeks to a month, the experimentees were given a survey (found in Appendix B) in order to get some qualitative data from their experience with the decision assistant and their given variant.

3.3.4 Phase Four: Results

The last phase of this master thesis incorporated the collection of the results from the experiments, which provided the evidence needed in order to assess the research questions. With the analysis of the results from the experiments, showcased in 6.2, a framework was constructed and included in chapter 7.1.

3.4 Data Collection

In this thesis collection of data was carried out in two different phases, as mentioned in the previous section. Such phases were a case study and a set of experiments. The former phase consisted of the analysis of the the developed tool, which in turn was used to generate a series of A/B tests over sale processes. The latter phase involved the conducting of experiments with the variants created during the first part of the study. The data collection techniques utilized are going to be presented in this section.
For the case study, qualitative data was collected concerning both the desired metrics proposed by the participants using the tool and their impressions after using it. The selection of the appropriate subjects for this phase was decided in a meeting with the supervisor of the Thesis in the company, who, based on their suitability, proposed a few candidates. This selection was further expanded during a general meeting of the company in which the authors presented the developed software. After this presentation, a brainstorming session was hosted in which participants were asked to provide ideas and target groups for the experiments. As a result of this general meeting, the initial selection of candidates was expanded to a number of no less than six employees interested in participating in the case study. Furthermore, one of the managers offered himself to act as a contact person or gatekeeper [41] during the interviewing process, ensuring that all participants were informed and coordinating the different schedules and interviews. The company supervisor agreed to act himself as a participant of the case study as well. In order to collect data, semi-structured interviews were conducted with all the participants. The main idea behind them was to identify which are in their opinion the metrics that will improve the success of the sales process proposed in the tool. Based on those metrics, and for each of the interviews conducted, the different systems were created -in both control and variant versions-, and published online.
The interviews were conducted in accordance to guidelines proposed by Runeson and Höst [42]. Every session started with a semi-structured approach with very few questions predetermined. The semi-structured protocol strengthens the exploratory nature of the study. According to the outcomes on those questions the discussion was expanded to gather more feedback. Following these recommendations, and after asking for permission from each individual interviewee, each session was recorded in an audio format, since even though one of the interviewers was focusing on taking notes, it is still hard to note down all details. As the interviews were not the main part of the study but were rather conducted with the idea of gathering information to set up the experimental phase, the authors estimated in accordance with the thesis supervisor that a number of around five interviews would suffice.

There was a need to identify the link between conducting both a case study and an experiment. Citing Runeson, "Case studies do not generate the same results on e.g. causal relationships as controlled experiments do, but they provide deeper understanding of the phenomena under study. As they are different from analytical and controlled empirical studies, case studies have been criticized for being of less value, impossible to generalize from, being biased by researchers etc". Case studies are useful in presenting different kinds of evidence and produce strong and relevant conclusions, but will never provide statistical significance due to the nature of the data collected (quantitative) [42]. On the other hand, experiments can be used to reinforce, accept, or even reject the previously conceived ideas, so a live experiment is seen as an appropriate way of providing feedback regarding these ideas [2].
After performing the case study it was time to move on to hosting the experiments. Since the previously conducted case study revealed a set of metrics, preferences and opinions regarding what could improve an online sales process, the developed tool was used to create a series of A/B tests which were accessed online by a selected sample consisting of real contacts and customers of the company. As in the selection of potential interviewees, the supervisor took a key role in choosing the experimental sample. The goal behind conducting experiments was to analyze if the conclusions extracted on the case study matched real life tests performed by customers, and not rely solely on the information provided by people from the company [3] [2].
After performing this phase, the authors were able to address the first research question, that is, how can A/B testing be extended from visual aspects of online services in order to optimize sales processes in the E-commerce domain. By setting up the experiments, analyzing its outcomes, and comparing those results with the interviews conducted, the authors are able to answer this question. In this sense, the experiments were useful to support the findings obtained from the interviews.
The experiments were set to collect both qualitative and quantitative data, although the key part of this phase was the quantitative data. In order to capture the user behaviour, the tool implements a tracking system which will follow the user during the course of his interaction with the website provided. By using the tool and its tracking system, qualitative data will be gathered using surveys displayed to the participants, while in the event of collecting quantitative data as well, different tools may be used such as scripts, logs, or other available methods.

3.5 Data Analysis

As mentioned in the data collection section, this thesis will comprise gathering both qualitative and quantitative data, extracted from a series of experiments and a case study. As such, following the division of phases established beforehand, the analysis of this collected data is presented in this section. The objective of analysis is to extract conclusions from the data collected. In order to do this, it is necessary to keep a systematic procedure in which the reader can follow the derivation of conclusions from the data [43].
During the case study, the collected information from the interviews comes in the form of notes taken during each meeting and its correspondent audio recordings. Following the guidelines of Runeson and Höst [42], the analysis of this information

was performed at the same time as the collection. This reinforces the idea of case studies to be flexible research methods, where new insights can be discovered and explored in subsequent iterations.
In order to reduce bias, the authors took an open perspective into analyzing the collected data from the interviews. This meant trying to abstract preconceptions extracted from the main hypotheses of the thesis and disregard them while trying to derive conclusions. The fact that both authors participated in the case study analysis, first in an individual basis, and combining their findings in a second iteration also helped reducing bias [42].
This approach led the authors to perform what Runeson describes as hypothesis generation techniques, in which the main idea relies in finding hypotheses from the extracted data. The decision to use this kind of techniques relies in the exploratory nature of the study, which tries to find acceptable criteria and metrics for the automated sales tool [44].
Once hypotheses are generated on which metrics and parameters would affect an online sales process, the experiments took an important part into giving statistical significance to those hypotheses. These experiments were created by selecting some of the hypotheses generated after the analysis of the collected data from the case study. The authors built up a series of experiments on the Decision Assistant tool with the assistance of some employees of the company. The experiments had a target group related to those characteristics, and a timeframe for the conduction of each one of them was set.
Once the timeframe set for the tool to be active online expired, the authors set the appropriate statistical tests concerning each null hypothesis. The variables and ranges set for each test are detailed in 6.1, as they are related to the different metrics set. However, following [2], confidence level and power were set respectively to 95% and 80-95%. More information on that can be found in 6.1.
Success or failure of the experiments lead to either a positive or a negative view in whether the metrics obtained and set during the study are useful for extending A/B testing, using a tool such as the proposed one, for the domain of E-commerce. Essentially, this analysis accepts or rejects the null hypothesis that can prove whether the A/B variants have any statistical difference. The collected qualitative data also contributes to reinforce these results. In addition, the metrics proposed in the case study will also be studied to evaluate their support over the results of the experiments. In case of the first research question becoming accepted, these metrics and the result of the experiments are in turn going to be the basis for the proposed framework addressing the second research question.

3.6 Threats To Validity

One critical limitation for this research applies to its construct validity, specifically the way the experiment is designed. Construct validity reflects to what extent the research methodology captures studied concepts and what is investigated according to the research questions [42]. As the research aims to provide conclusions based on quantitative data, the need for a sufficient sample size is essential. However, this problem is compensated by the fact that the experiments will be conducted over real

samples provided by the company with approximately one hundred persons each. Also to reduce bias during the selection of subjects a "gatekeeper" at the company was used. That is the authors of this thesis did not influence the selection of subjects. Another threat to construct validity relies in the nature of the experiments. Double-blind experiments indicate that researchers themselves can be threats to construct validity, and studies should attempt to control this effect.

This is linked to another potential threat. The authors identified that construct validity could also be threatened from the inclusion of the Thesis supervisor at the company, the CEO of the company, in the interview process since from his position as a product owner he could skew the results for his own benefit. To counter any such allegation the authors of this thesis would like to state here that his capacity as a supervisor was merely communicational towards the rest of the company's employees and his involvement in the development process was only for creating the general idea and the validation of the developed tool. His involvement in the interview phase was done from the viewpoint of an established sales professional rather than from that of a product owner. That being said the results gathered from the interview process could not have been skewed since he could not have any gain from biasing his answers. This is further reinforced by the fact that the characteristics selected for the experiments were not based on a particular interviewee point of view, but rather from the compilation of all ideas and the generation of hypothesis based on that compilation.

Another potential threat can be found on internal validity. Internal validity refers to the risk of interference in causal relations within the research. For instance, if the researcher investigates whether factor A affects factor B there is a risk that factor B is also affected by factor C. Researchers need to pay close attention to alternative factors that might affect the investigated factor, otherwise there is a threat to internal validity [42]. Since the first part of the study is going to be performed cooperating with at least five or six employees of the company, which will be creating the different website variants necessary for the experiment, there is a threat of them manipulating the variants of the web sites so that the experiment will throw the results they personally aim for, instead of real business objectives.

The authors account also for a potential threat that can also be found with regards to the external validity of the research [42]. Specifically to what extend can the findings be generalized in order to produce a suitable answer for the second research question. This can be aleviated by taking into consideration that the company in which the experiments took place cooperates with other companies which would allow the experiments to have a much more wider target group than just that of a single company.

Finally, as Kohavi [3] also states, choosing the right metrics for the experiment is a qualitative decision and can in the worst case lead to detrimental outcomes when selecting the wrong metrics. This is connected to limitations in hypothesis generation. This technique, which is used during the data analysis of the case study constitutes a challenge in which the researchers must show innovative thinking and a strong ability to analyze and generalize findings from qualitative data. In addition to this, Kohavi [2] points out that while carrying split tests it is possible to know which variant is better, but not the reason why it is better. This limitation can

be solved by additionally collecting users' comments. This study addresses this limitation by providing a short questionnaire to the experiment subjects, in order to complement the experiments.

3.7 Main Contribution

The objective of those questions is to improve the business value of the system and ultimately lead towards a generalization of A/B Testing in such kind of systems. This generalization will be expressed in the form of a framework, consisting of a set of guidelines, that can be exploited in similar environments.
The knowledge gap this thesis intends to cover fits with the description that Hynninen and Kauppinen [34] cite. They propose that "A/B testing can complement qualitative user research and offer a potential way to validate the value which system improvements bring to customers". In this research, the authors contribute to that statement by performing the cited combination of A/B testing and qualitative user research, in the case of this study performed by sending a small questionnaire to the experiment subjects upon completion of their interaction with the website.
This thesis also contributes on the research field by adding to the importance of supporting A/B testing in the E-commerce domain by escaping the norm of testing usability aspects (e.g. creating variations of the GUI of the site, such as different layouts, fonts, colors, etc) and instead perform tests over aspects related to business processes, which in this particular case is the sales process.

4

The Decision Assistant Tool

The following section presents a description of the produced tool that the authors of this master thesis developed and used in order to carry out the case study and the experiments.

4.1 Purpose & Characteristics

The Decision Assistant tool was created in order to help businesses promote their products or services by assisting their website visitors in evaluating whether those could be useful to them or not. The Decision Assistant tool has two potential types of users. The first type of user is the tool administrator. The administrator belongs to a business that wants to sell a product or a service, and is responsible for creating a Decision Assistant Flow. The other type of user can be described as an end-user or target user. This end-user can either be a business or a private person, and its goal is to go through the Decision Assistant Flow in order to get information about a product or service and with the help of the Decision Assistant tool come to a decision on whether this product or service fits its necessities. In what can be described as a (but not limited to) survey-like creator tool, the administrator (business owner) can create a flow of steps, called a Decision Assistant Flow or DA Flow. This flow can provide information to the end users by using a wide array of items, such as images, videos, HTML formatted text, and diverse types of questions. Among these questions, the Decision Assistant supports the following:

- **Open questions** ask for information on a general topic. They do not have any predefined answer, instead the user inputs text to describe either a problem, situation or as a way of giving feedback.
- **Single choice** questions have a predefined set of answers of which the user must choose only one.
- **Multiple choice** questions are presented in a similar way as single choice, but with the particularity that the user is allowed to choose more than one answer.

The purpose of the flow can be to help the visitor of the website (the previously described end-user) ascertain whether the proposed product or service is exactly what he is looking for, or in its simplest use, to just inform the visitor about the product or service. The administrator of the flow can also assign a weight to each possible answer. Once the visitor enters the flow and advances through it, the

system keeps track of the selected answers, in order to conduct calculations with the weights. By the end of the flow the weight of each selected answer can be added up, and be compared to a threshold value defined by the administrator in order to provide the visitor with either a positive or a negative suggestion.

Each step of the sales process can be represented within a Decision Assistant flow. It can either take an unaware visitor and inform him on the specific product/service, or take a business contact and turn him into a client by successfully calculating his needs and suitability. The flow can start by making an introduction of the product or service, presenting some general information on its field of use and also the advantages that it can give to a potential owner. The visitors that decide to follow the flow can then be taken through a number of steps containing questions designed to assist him in order to gain any kind of information he needs. A demonstration of the product can follow, either through a video presentation or images and text. By the end of the flow the visitor can learn whether his answers indicate that he would gain from using the product/service or not and also be required to enter his personal information in order to continue, or get more personalized feedback by email. The flow can also contain an order form in order to eliminate the need of any redirections if the visitor is about to make a purchase. It can be also utilized to represent the next steps of support and feedback as it can save the visitor ID along with his current step of the flow for his later visits at the Decision Assistant, and is set to include future expansions, like for example a live chat tool in order to better facilitate the support step.

4.2 The Architecture

The Decision Assistant tool was designed using the Model-View-Controller (MVC) architecture. The Model part of the architecture includes a database module which contains all the building blocks needed to create the Decision Assistant tool. The database module is wrapped by another module which has the task of connecting the database table to their model representations in the back-end of the tool. Any read/write operations on the database is going through that module. The Controller part of the architecture includes a module that's responsible for manipulating the models and passing the information to the graphical user interface. Also in the Controller there is a module responsible for handling the A/B tests, and one responsible for rendering the Decision Assistant flows. The View part of the architecture includes the two variations of the graphical user interface: the administrator side (flow editing tool) and the end-user side (flow display tool). The bulk of the development was made using the open source Laravel PHP framework (Controller). MySQL was used for the database (Model) and HTML5 along with jQuery, CSS3 and Bootstrap was used for the front-end of the web application (View). As wel as those, several minor open source JavaScript libraries were used mostly for cosmetic reasons on the front-end side of the web application. The layered view of the application along with the tools used to achieve it can be seen in the following diagram:

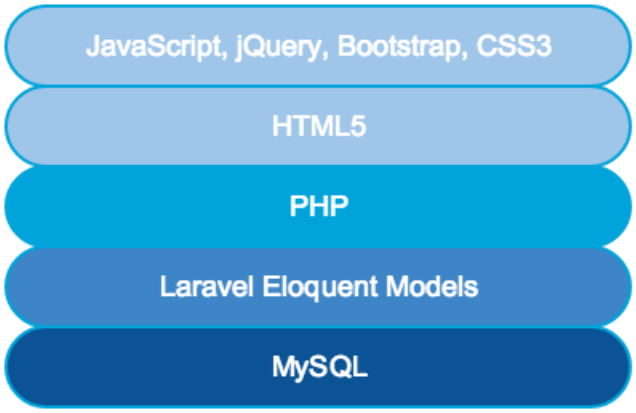

Figure 4.1: A layered view of the used tools ranging from back-end (darker) to front-end (lighter).

4.3 A/B Testing Support

The Decision Assistant tool offers also A/B testing capabilities for the sales process part of the flow. It can range from variant sequence of questions and/or steps to status aids for the visitors. When the administrator is ready with the design of his Decision Assistant flow he has the option of cloning the whole flow and modifying this clone in order to produce the second variant (Treatment variant). The modifications available are:

1. Reordering, joining or breaking down the steps of the flow.
2. Reordering the questions of each step or moving the questions to different steps.
3. Altering the weights of the answers or the decision threshold,
4. Providing different routes for the flow via specific answers (skipping steps for example)
5. Adding a progress bar to let the visitor know the completion percentage of the flow.
6. Allowing the visitor to see the specific location (current step) he is at and the upcoming (next step).
7. Allowing the visitor to see the total length of the flow along with his current position (number of steps - current step).

After choosing from one or more of the above possible modifications the newly created clone will be connected to the hash url of the Control variant. Whenever there is a new visitor, one of the two variants of the flow is presented to them randomly with 50% probability. Recurring visitors will always get the same variant they were first assigned to (provided of course they allow cookies or use the same

system and browser to view the flow website). The results of each variant can be then reviewed individually and be compared in order to define which of the two can be more successful (in terms of successful outcomes), and help guide the administrator in the design tactics for his future Decision Assistant flows.

5

The Case Study

The following section presents the case study that was conducted during this thesis. The interview process is presented in the first place, followed by the analysis of the interviews that were hosted.

5.1 Interview Phase

After the development of the Decision Assistant tool the authors along with the members of the company for which it was developed held a brainstorming session in order to come up with possible experiments which could be instantiated using the tool. Then, each member was interviewed in order to gain insight on their ideas and thoughts about the tool's usefulness as well as the formation of the metrics and configuration characteristics to be used in the coming experiments. During the case study and the experiments the tool was used under real conditions. In total seven employees or partners of the company participated in the case study.

5.1.1 Background

The selection of the appropriate subjects for this phase was decided in a meeting with the supervisor of the Thesis in the company, who, based on their suitability, proposed a few candidates. This selection was further expanded during a general meeting of the company in which the authors presented the Decision Assistant tool. After the presentation, a brainstorming session took place during which participants were asked to provide ideas and target groups for the possible experiments. As a result of this general meeting, the initial selection of candidates was expanded to seven employees or affiliates that were interested in participating in the case study. Furthermore, one of the managers offered himself to act as a contact person or gatekeeper [41] during the interviewing process, ensuring that all participants were informed and coordinating the different schedules and interviews. The company supervisor agreed to act himself as a participant of the case study as well, but from a merely sales expert perspective. The participation of the company supervisor could have enticed some risks on whether his points were biased into directing the case study towards personal aims or goals for his specific business, something that the authors had to aoid. The seven case study subjects were the CEO and founder of the company, the Chief Of Operations, the Managing Director and the Delivery Manager as well as two company advisors and a partner.

Interviewee	Role
Number 1	CEO & founder
Number 2	Managing Director
Number 3	Chief of Operations
Number 4	Delivery Manager
Number 5	Partner
Number 6	Advisor
Number 7	Advisor

Table 5.1: The interviewees and their role in the company

5.1.2 The Interviews And Their Purpose

Over the course of a week the case study subjects underwent individual semi-structured interviews with the authors. The purpose of the interviews was to validate the Decision Assistant tool and whether it could be of help to the company. Also to gain insight on the important factors when it comes to creating an online sales process and based on that create the Decision Assistant Flows that were used in the experiment phase.
The interviews were conducted in accordance to guidelines proposed by Runeson and Höst [42]. Every session started with a semi-structured approach with seven predetermined questions. According to the answers on those questions the discussion was expanded to gather more feedback. The interviews lasted about 25 minutes on average, with the longest lasting 45 minutes and the shortest 15 minutes. The whole question sheet can be found in appendix A. The following subchapter contains more information about the purpose and the goals behind each question.

5.1.3 The Questions (Goals & Purpose)

The main idea behind the questions was to identify which are, according to the interviewees' opinion, the metrics that will improve the success rate of the online sales process proposed in the tool. Based on those metrics, and for each of the interviews conducted, the different flows were to be created, in both control and variant versions, and published online.
The first two questions of the interview had the goal of facilitating the interviewees and creating a bond with the interviewer. The subjects were asked to give information about their background and position in the company. Also since the interviews took place one week after the presentation of the Decision Assistant tool and the brainstorming session, they were asked about how many details did they know about the tool before the meeting and whether they tried to get more acquainted with it before the interview. This question also helped weed out those not that interested with the experiment and identify those that were more invested in them.
The next three questions were regarding their opinion on a successful sales process strategy and how the Decision Assistant tool and its' customizations could help them fulfil those. The obvious purpose of those were to gather information on the important metrics the subjects would like to be used for their experiments.

Additionally those questions served also as some form of validation for the Decision Assistant tool, since if the subjects could see ways in which the tool could help them in their sales processes that would add to the tool's usability and its' value.
The last two questions were about the target group for the experiments as well as for the decisive differences between the two variants. Those questions were of a merely informative nature with the purpose of getting a preview on the expected samples for the experiments and the experimental Decision Assistant flows they would be exposed to.

5.2 Results Of The Case Study

The case study's main technique used to gather data were the interviews conducted at the company's offices. Those interviews focused on the validation of the Decision Assistant tool and on gathering information about the participants' ideas on how to set up a successful online sales process. More on how to achieve success in an online sales process is explained in the following subsection. The interviews were also designed with the objective of also gaining insight on the metrics and/or characteristics that are deemed as important in order to be used in the experiments that followed.

5.2.1 Overview

The Decision Assistant tool developed by the authors of this thesis received high praises from the company personnel and partners. Something highlighted in their responses was their certainty that the Decision Assistant tool could help them achieve their goals and their eagerness to test it in real time environments. Most of them had either heard of or had some experience with it before the interviews and thus they could actually verify that they tried and achieved what they wanted to do. A successful online process is, according to most of the interviewees, a process that is able to transmit to the customer a believable profit or benefit when the customer is looking for a product or a service. It must show this benefit in a clear way so that the purchasing decision is done without doubts by the customer. In turn, from the seller's perspective, a successful process is one that leads to a comparatively good number of conversions (purchases, acceptances). A successful process also must be tailored to the needs of each customer or customer group. The Decision Assistant tool can support these beliefs through its' customization features and its' adaptability to any kind of sales environment. Most of the interviewed subjects were also looking forward to test many a sales method during the experiment phase. In the following subsection a thematic map which was generated from the interviews is presented, and after it some more details follow along with quotes from the interviews.

5.2.2 Thematic Map

The analysis of the data gathered in the interviews showed two main themes, namely Successful sales processes and A/B Testing characteristics. Those themes provided a set of codes that can be found in Table 4.2, and further described in the following

sections 4.3.3 and 4.3.5 respectively. Section 4.3.4 shows another theme treated in the interviews, the Decision Assistant tool validation. However, given the different nature of this topic, it was decided not to be included in the thematic map.

Global Theme	Organizing Themes	Codes
Extension of A/B Testing into E-commerce environments and online sales processes	Successful sales process	Believable benefit Easy next step Information Professional website References Capture leads Strategy Trust
	A/B Testing characteristics	Irrelevant questions Process Progress bar Reorder questions Different paths Length of the process Poking questions Language variations

Table 5.2: Thematic map

5.2.3 Successful Sales Process

In order to have a successful online sales process the following codes were observed in the interviews: A believable benefit for the website visitor, an easy next step, reliable information, a professional looking website, relevant references, and a good preparation involving lead capturing, a good strategy and building up trust between the seller and the customer.

Believable Benefit
A believable benefit would mean that the visitor of the website can get something either for free or for a bargain price by buying the product that is offered. This believable benefit could also be tailored for each specific visitor. The importance of such a thing is something highlighted by three interviewees with N1 saying: *"the only thing that you need to do is provide a believable benefit [...] particularly for the recipients, and even specific to the recipient."* N5 saying: *"I believe it's very important for the buyer to easily [have an] insight of the benefit for the company or the service then it's gonna be easy"* and N7 saying: *"finally i would say get some kind of free [bonus] opportunities".*

Easy Next Step
Ease of use during the online sales process is also a very important factor. The visitor must be able to easily find his way through the order forms and product/service information so that he can take the next step without much confusion. This factor

was noted by two of the interviewees with N1 saying: *"[...] And then, an easily identifiable call to action for the next step"*, and N5 saying: *"[...] an easy way to take the next step (trial period etc.)"*.

Information
Just as the lead capturing system can lead the visitor to the information he wants it's equally important that this information exists and is of a certain standard that is easily understandable and relevant. Answering the question from the website visitors point of view N3 remarked: *"something [that] I want [is] information, I want background, I want to know what I'm buying [...] I think that it is important that the customer, the buyer, gets a feeling of what it is all about, and information and conviction. Information I think it's the most important."*.

Professional Website
A professional looking website is obviously of the utmost important for a successful online sale. As N7 highlights: *"I think especially with things like this, [a] very professional website, great landing pages, great digital content, and [really] gain thought leadership in the specific area that you find yourself."*.

References
A professional looking website can never exist without showing user testimonies and references from well known persons or organizations that use the the service that is on sale. N6 states about this factor: *"The next thing that's very important is to have some references so you can identify yourself with other different companies or positions"*.

Capture Leads
Lead capturing is the process of pointing website visitors in the right direction when the want to get more information. In order of a online sales process to even happen the seller must have implemented a good lead capturing system. Being able to achieve that can lead to a successful online sale. Two of the interviewees mentioned this specific factor with N6 saying: *"The key factor is actually to capture leads. As soon as visitors reach the website it's important to capture them [and] in a way make them interested in your product, your services or whatever [you want to sell]"*, and N7 saying: *"[The seller needs to have a] clear understanding and common ground on alignment with lead scoring, [know] what is a marketing qualified lead, what is a sales qualified lead, [and] how to have a very, very effective online leads source; leads source is basically how you get your leads"*.

Strategy
Before initiating the online sales process a factor that could lead to its success is the strategy of the seller. The plan of action must be decided beforehand. The Seller must have a very good and clear understanding of all the variables that involve his product and his target audience. Two interviewees pressed for the importance of this factor with N2 saying: *"A good strategy. A good understanding of what you need to do, and [how] to implement it. The system is not so important."*, and N7 saying: *"[...] understanding the market and segmenting it to your best ability and I'm constantly underlining the importance of customer profiling. [...] Profiling and really understanding [the visitor's] decision and buying process, really understanding that"*.

Trust
Being able to achieve a certain level of trust with the website visitor is also something required for a successful online sales process. As N4 remarks: *"Trust. Since the Internet is such a big place, where there are lots of companies in the same field [...] and all want to do the same deal with their customers. So in the end what the customers are deciding, the thoughts of [whether] they want to go with one company or the other, it [depends on] how trustworthy is the company. So I think trust weights in a lot."*, with N1 coming to partly agree but also highlight that focusing on building up trust is not the only thing an online seller must do: *"[...] In short, to be trustful, offer as much transparency on your side as possible, [although] it's not enough."*.

5.2.4 A/B Testing Characteristics - Metrics

During the interviews, the interviewees were asked to give their opinion on which of the adjustable characteristics of the Decision Assistant tool would come in handy for them. Upon those would the experiment flows be based. The following characteristics and possible variations were highlighted by the interviewees.

Irrelevant Questions
Irrelevant questions are defined as those which do not totally belong to the purpose of the process. Under this category are questions which ask for wrong information, information that is not connected to the process, questions that will lead into acquiring useless information from the visitors of the process or even questions that might ask for personal information that visitors could be unwilling to provide. As stated by three interviewees, the relevancy of the questions ranks in the first position among the characteristics. Interviewee N1 states: *"If you would have irrelevant questions that would be clearly irrelevant even to the concept itself it feels like you normally wouldn't do...would really try to avoid irrelevant questions but perhaps for an A/B testing scenario this might be useful [...] If you think of it irrelevant questions could be one thing, but even something that would be completely logically wrong, if you encourage people to tell you and come with feedback, would they give you feedback on things that they would understand it is completely wrong or not?"*. This interviewee explains that having irrelevant questions, while being a logically wrong idea in the conduction of an online sales process, might be of use for an A/B Test on this field. Interviewee N3 also considers this a possibility: *"Is the question too relevant? So I can see if it's an effective test."* To wrap up with the relevancy of the questions, interviewee N6 focuses on this aspect in having the right amount of questions. N6 even goes beyond into hinting that too personal questions could force people to quit the process, since the goal should be to show off the product instead of trying to obtain personal information. *"I think it's important to not have many questions and not too much details because you will, there are people that will stop answering questions. It's important not to have many questions and try to have the most important [...] otherwise they might get bored or just irritated that you are trying to figure out so much about them instead of you should make them interested in your project, so you want them to discover your product instead of you should have questions about who they are cause then the risk is that they will not do that."*

Process

Interviewees N5 and N2 agree that appropriately constructing the sales process is the most important thing, ahead of the tool or method used. N5 states: *"It is always a process that is the most important. [...] Business process is vital and its should be rated at the top."* and N2: *"But just having the tool I think it's not enough."*

Progress bar
The Decision Assistant tool comes with the option of including a progress bar for its processes. The function of the progress bar is to guide the user and give him/her feedback on the current progress they have achieved in completing the process. The discussion of whether this would encourage more users to finish the current process was arisen by two out of seven interviewees. Interviewee N2 states that this is one of the most important functions, as shown here: *"I think that the tracking function that you can see where are the the test-doers, what stage is he. I think that one is pretty much one of the most important functions, that you can see where does the testing end [...] It is quite important to have, to see, to estimate the time actually. Because if I'm doing a tests, for like 15 minutes, and I see I've done only 20% of the test, I may not continue. It's important to estimate the time, I think that the progress bar is one of the primarily functions."* Also, interviewee N5 considers this important with the following reasoning: *"The buyer or the person that you're having this dialog with should be aware of where he/she is at this moment."*

Reorder questions
Another functionality that the Tool incorporates is the possibility of changing the order of the questions in the variants, or even to completely edit them. Interviewee N3's opinion is that different orders for different people might reduce the possibility of having a boring process: *"If you scramble the questions, and people get the questions in different orders, I think that way you avoid the boring logic of somebody else doing the test."* As well as N1, which has the opinion that these two options are of high importance: *"In a way I guess the edit and reordering will be helpful, because it might be difficult to come up with the right scenario directly [...]. So the reordering and the of course editing is going to be quite important."*

Different paths
The possibility of altering the path that the user goes through in the process was also identified as a potential factor. This is a functionality provided by the Tool which allows a user to go through different stages during the process, and which is the next stage is decided by the tool based on the input the user has been giving through the previous step(s). Interviewee N4 regards the following about the possibility of offering different paths in a process: *"If they take step A then they should get to step B but if they take different answers from step B they are taken on to different paths, and that's really a huge thing"* and to the question on whether this would be useful, N2 states the following: *"I think so."*

Length of the process
The length of the process is related to the previous characteristic of the relevancy of the questions and the progress bar. Having a too long process might provoke a user to drop out, and this can be even be aggravated by the fact of combining a too long process with irrelevant questions or not giving the user feedback on his progress. This aspect of a lengthy and non-specific process is stated by N7 as follows: *"If it's*

getting too long you're gonna bore people to death, so really really getting the specific questions and what are those questions that is for me the critical part."

'Poking' questions
Interviewee N1 identified that perhaps some users might not be willing to answer some "agitating" or "poking" questions. This interviewee identified this with a practice that sometimes happens in sales, in which a salesman tries to find a problem and make the customers understand better a possible solution. This characteristic has to do entirely with the formulation of the questions and does not rely on a specific functionality provided by the tool. N1 states the problem as follows: *"When I think about it, one thing that could be interesting to have is 'quite poking questions' [...] to see if it the problem can be agitated more, then again, that's a little bit of a risky thing, it will most likely make someone to think more about their potential problem and also to understand more about the possible solution that would be a benefit to them. [...] My personality type is not the one that I would want to poke too much into someone, whereas others may do it even in sales situation and they are probably more successful with it, [...] quite interesting to see if there would be a step that would have really poking questions, and then perhaps not include that step in another variant, and then see: do people actually answer those very poking steps or not?"*

Language variations
The main issue with communication is to know who is the target audience, with the objective of using an appropriate language. Interviewee N4 proposes an example of this fact, and states it as a possible fact which could be taken into account when performing an A/B Test over an online sales process: *"I can simply do a test in where I ask questions in plain simple English, and then another with more like rasta style, so more or less, a variation in how are we actually communicating with our customers. That's the most basic thing in communication, knowing your target audience, how you are going to communicate with them."*

5.2.5 Decision Assistant Validation

The validation of the Decision Assistant tool was thought to be key to the success of this research. The authors define "validation of the DA" as the capability or suitability of the tool to conduct sales processes and help both the seller (the provider of the goods or services) and the customer into better understanding their needs. This was addressed by all the subjects of the Case Study in the conducted interviews. Upon the question "How do you think the DA is going to help you achieving that? (A)", here we present the results obtained after analyzing the answers of the interviewed candidates:

Interviewee N1
This interviewee claims that the DA Tool would indeed be capable of conducting successful sales processes and explores the inherent psychology of the tool: *"I really have really great hopes about that it will do exactly these things, I think the suggestion of process will make it more believable. It will be easier to convey a benefit, such*

as a person will understand, can also do it [achieve the benefit] with questions, [...] I'm thinking now more psychological perspective."

Interviewee N2
This interviewee starts by giving first a partial approval: *"Yes, of course. It's a perfect tool for implementing [a] good strategy inside. But just having the tool I think it's not enough."* After this, the interviewee continues his reflections on the strong points of the tool, which according to his/her beliefs are the easiness with which the tool can present information to customers and present them with a benefit: *"the new way of doing it, because you can do it with a banner ad, or with content mail or whatever, but the picture is very valuable for the customer. I can really see a need for it, and it's very easy to present benefits for me."*

Interviewee N3
This interviewee considers the fact that the main question drawn by a potential customer is related to how can a product or service solve his or hers problems. His quote is presented here: *"I can see the DA like that, because you get to know about the process and about everything, I think that you get the picture of "how does this suit me?"*

Interviewee N4
The idea that interviewee N4 wants to transmit is that the tool can be validated, and this interviewee puts a bigger emphasis on the customization possibilities, concluding with the idea that although the tool might look like a simple survey, it is actually much more complex and powerful than it seems. *"Yes, I can see many uses, for example, In the [promotional flow made by the company] there are a lot of things that we can do [...] to make it very custom, and also showing the customers if they [think they] are doing a survey, it is not like you create a questionnaire and answer this and answer that, they can set it in so many instances"*

Interviewee N5
This interviewee found out the tool to be quite suitable for his needs in marketing automation and sales by pointing out the huge amount of tools available, and his/her thoughts that the tool will be helpful in narrowing down the real needs of people involved in E-commerce and online marketing, but also for customers to see the real benefit that they are looking for: *"I believe it is gonna help us much because we're really contacting companies and people regarding marketing automation and we talk with several marketing directors or sales directors. Especially marketing people are overflown with tools or systems that can help them, you know, there are hundreds of them, so to be able to nail down and really work through what they really need now and to help themselves in their buying process, because it is so much to choose. And since it could help me in my sales process the customer can more accurately see, oh this is what we need at this moment."*

Interviewee N6
Interviewee N6 starts by pointing out, as many other interviewees, that the Decision Assistant can clearly show a benefit to the customer: *"I see that the decision assistant could be of help actually, to make them [the customer] understand that perhaps [we] have something [...] which can help them now or in the future"*, *"I think it is a*

good way to guide the visitors, and help them understand what this is all about, what marketing automation all about, and with asking certain questions which go deeper and deeper it could help them understand [...], so I think it is a hand to explain to the customers [...] understand they are in a need of your service or product. It is help to guide to make a decision, to guide them and help get an understanding of what they might have." The last remarks of interviewee N6 are related to those of interviewee N5 in the sense of being able to apply the system for his/her own business: *"Actually, when i saw your system for the first time, the D.A., my first thought was that I could have use of it in my own business to gain more prospects and like that,"*

Interviewee N7

This interviewee claims a similar argument as N5 and N6, expanding it from his personal needs to what he believes the customer can see the benefits of conducting this kind of sales processes. *"It would help me really as a sales enablement technology, And that person basically needs to make a buy decision, and you are really in the conversion stage, you want to convert him, that's where i think definitely the tool will help our purpose."*

6
The Experiments

This section reports and discusses the setup and results of the experimentation conducted. Section 6.1 presents the setup of the experiments. It includes an overview and reasoning of the idea of hosting experiments, as well as details of the characteristics selected for each individual experiment. Section 6.2 presents the results of the experiments.

6.1 Setup

The idea of hosting experiments was conceived in order to analyze whether the thoughts and ideas of the interviewees about A/B testing in e-commerce were right, or at least were in the correct direction. They were hosted right after the case study was performed, since it was necessary to at least partially analyze the interviews conducted with the goal of setting up the tool with the correct parameters. The outcomes of the experiments were expected to help reinforce the answer to the first research question of this study.

6.1.1 Overview

Out of the seven interviewees from the case study reported in Chapter 4, three of them agreed to get involved and test their ideas in the experiments. Due to time issues though, one of them had to pull out of the process. In order to have a better chance of getting positive results from the experiments the authors of this thesis decided along with their supervisor to create a flow of their own to test. The three experiments took place during the course of weeks 31 through 38 of 2015. The timespan set for each experiment was limited to two weeks. It was decided however that the duration of an experiment could be shortened or prolonged based on the amount of responses received.
Each participant created a flow using the Decision Assistant tool, in which they defined a control and a treatment variant. The differences in those variants were based on the thoughts and ideas they provided during the interviews in the case study, which were in turn analyzed in order to generate hypotheses which served as the base of the experimentation. Once each of them created the flows, a target group was selected among contacts of the company, and the flows were sent to them via an e-mail.
Based on Kohavi [3], the tests were carried out anonymously. This means that the subjects of the experiments did not know that they were being part of an online

experiment. As part of the development of the tool, the authors created a script which acted as a load balancer, which distributed all participants randomly to one of the variants. Once the participants entered the flow, there were two possible outcomes: finish the flow or drop-out at some point. Since the criteria to consider an experiment successful was for the participant to finish it, special care was taken in order to track the participant's status regarding completion of the flow. In order to achieve this, the tool would provide continuous tracking information to the authors, stating how many different participants started the flow and their status (seen, doing, or finished). In case of the 'doing' status, there might be two different possibilities; that the participant has stopped completing it but intends to resume it later, or that the participant decided to drop out and will not complete the flow.
The authors also prepared a very short survey intended to obtain further feedback from the subjects (the target group) of the experiment. The survey was sent to each experiment participant approximately two to three weeks after the first send out. The reason was to get some insight on the cause for their completion or dropout of the flow, and suggestions on how to improve the process and the tool.
In the following section the setup for each experiment is described. This includes the motivation for conducting it, which is related to the hypothesis to test. This motivation describes the characteristics of the control and the treatment variants, the size and characteristics of the target group and the timespan for the duration of the experiment. After that further information about the survey will be revealed. Results for each individual experiment can be found in section 6.2.

6.1.2 Experiment 1

This experiment was created by interviewee Number 7. The purpose of the experiment was to help the interviewee's company calculate its' client base's suitability for their product. The hypothesis that the authors extracted from the interviews and presented to interviewee Number 7 relies in testing whether or not having variants with longer processes and less dense steps differs to variants composed by fewer but more dense steps. Since the target group was completely comprised by Swedish speakers the flows were created in Swedish in order to facilitate their interaction with the flow.
Characteristics of the variants: The control variant consisted of 9 steps one of which was a finishing/thank you step. The first two steps had one question each while the rest consisted of 3 to 5 questions. Each answer had a specified weight with the end calculations leading to either a high suitability (success) or low suitability (failure). The visitors could get information about their status in the flow by a bar on the top of the screen which showed the percentage of the flow that was completed at their current step. The treatment variant was made by merging some of the steps together resulting to a flow with 4 question steps. That way the visitors that were exposed to the treatment variant would see a bigger completion progress whenever they moved to a new step.
Target group: The target group consisted of 8272 persons from the client-base of the Number 7's company.
Time span: The experiment started on Thursday 10th of September with the send-

out of the link to the target group. The experiment was deemed finished at Friday 18th of September.

6.1.3 Experiment 2

This experiment was created by the authors as it was deemed necessary to have at least three experiments in order to better reinforce the answer to the research questions. The purpose of this experiment was to promote and calculate the suitability of the authors' affiliates for beta testing a future product of theirs currently under development. The hypotheses extracted from the interviews was to test whether differences in levels of language formality actually have an impact among the target group.
Characteristics of the variants: Both variants of the flow consisted of 5 steps, one of which was a finishing/thank you step. For the Control variant the questions were asked in a formal manner, and only relevant information was inquired. In the Treatment variant the questions were formed in a more relaxed way, using also a bit of humor. In addition some questions were added to each step that were focused on gathering more private information. With those two variants of the flow the authors wanted to see the differences between using a formal language or a more casual one but with the addition of more intrusive questions (as per Number 1 remarks in 5.2.4).
Target group: The target group consisted of 612 people, friends or friends of friends of the authors. The age span of the participants fits the criteria for conducting an experiment with language variations. In addition, this target comes from different backgrounds, country of origin, and socio-economic environments and would under real life circumstances would be the target audience of a crowd-sourcing application such as the suggested product. Thus it can be characterized as representative for this specific study.
Time span: The experiment started on Thursday 10th of September, with the link of the flow posted in a social network. The experiment was deemed finished at Friday 18th of September.

6.1.4 Experiment 3

This experiment was created by interviewee Number 4. The purpose of the experiment was to reach out to the clients of the company and get information and feedback regarding their satisfaction with the company's support services. The hypothesis that the authors wanted to test relied in how visitors of a flow behave regarding the amount of imformation presented to them, such as showing them how far have they been in the experiment and how many steps do they have left to complete before finishing. Since the target group was completely comprised by Swedish speakers the flows were created in Swedish in order to facilitate their interaction with the flow.
Characteristics of the variants: Both variants of the flow consisted of seven steps, including an introductory and a finishing/thank you step. Each step contained a variety of question types, including an optional contact form. The difference in the

two variants was in the way the steps were displayed. The Control variant exposed the visitors to only the upcoming step which meant they had no idea on their exact position of the flow, while the Treatment variant exposed them to all of the steps which made them fully aware of their current position in the flow.
Target group: The target group consisted of 83 persons, all of which were clients of the company.
Time span: The experiment started on Thursday 30th of July with the send-out of the link to the target group. Due to low response it was resend at Tuesday 8th of September. The experiment was deemed finished at Friday 18th of September.

6.1.5 Surveys

Attached to each send-out, the authors included a short survey. The purpose of such survey was to further investigate the actions of each subject, especially when they decide to drop out during the course of completing the flow. The survey was the same for each experiment, and it included three short questions which inquired the subject on whether he or she had completed the flow or dropped out during the process, and in the event of drop-out what was the cause for it. Moreover, the survey also asked for extended feedback regarding either the tool or the flow presented. A copy of the survey question can be found in Appendix B.

6.2 Results

6.2.1 Overview

When it comes to the experiments, each one was examined individually through statistical analysis with the objective of reaching an understanding in whether there is a significant difference in the conversions among both the treatment and the control groups. In order to conduct this analysis, and since the values obtained from the experiments are categorical, the starting point will be the contingency tables created during the analysis of the experiments. These contingency tables present the figures for both outcomes of the experiment, success or fail, and the variant to which they belong. Afterwards, a Pearson's Chi Squared test for fitness will be conducted, in order to test whether there is statistical significance between the control and the treatment groups. Pure fails refers to participants who finished the flow and obtained a fail outcome, not accounting those participants who dropped-out. However, since the test is carried out using categorical variables, and the authors are only testing success and failed, dropouts will be added to those pure fails in the contingency tables created to account for total fails. The contingency tables will show absolute numbers, which refer to actual participants tested en each group, and not percentages, even though for better understanding of the reader, the authors will also use percentages to describe each experiment's result. Regarding whether the Yates's correction should be applied or not to a Chi-squared test, the decision was not to apply it, given the fact that both the sample size and the cell count were sufficiently large enough (cell count refers to the figure in each cell) in all

the useful experiments, and also to the fact that it tends to give very conservative p-values.

6.2.2 Experiment 1

The first experiment to detail corresponds to the one created by interviewee N7. As described in 6.1.2, this experiment targets real customers of the company in which N7 works at, and the sample size for this experiment totals 8272 contacts. The Decision Assistant tracking tool showed that 78 and 63 subjects were taken to Variant A and Variant B respectively. As specified in the description of the experiment (6.1.2) Variant B had fewer steps with comparatively more questions each than Variant A.
The results provided by the tracking tool of the Decision Assistant show that for Variant A 23% of the participants completed the flow, having a high rate of success. Thus, in total, 77% of the subjects dropped-out at the beginning or during the process, with 22% of the total finishers resulting as success and only 1% resulting as pure fails. Combined, the total number of failures sums up a 78% of participants. The relative figures for all finishes are then a 95% rate of successful finishes, with a 5% of failed finishes, although this figures are not accounted for the statistical test performed.

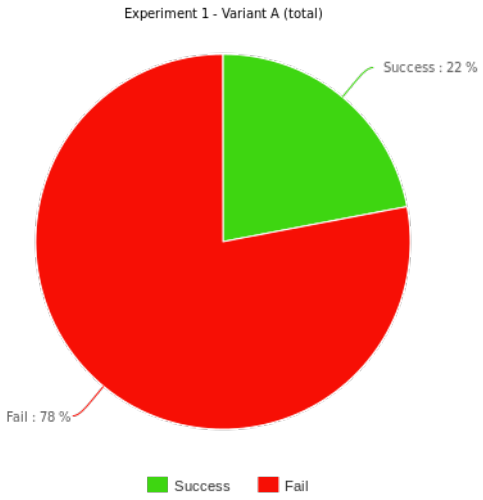

Figure 6.1: Experiment 1, Variant A Total (in %)

Variant B shows a lower 11% of finishes out of the total size, however with a 100% success rate among those, and consequently no rejections in this partial analysis. In total, this variant makes up for an 11% of conversions, with a total of 89% of participants achieving a fail status.
Regardless of the apparent success of Variant B over finished processes, the analysis is based on the total figures, including those who did not finish the flow as failed. Then, figures show a big difference among variants, indicating that shorter steps adequately classified and separated might make up for a more dynamic interaction with the system, and thus encouraging participants to stay and complete the process. With these values, a contingency table is presented as follows:

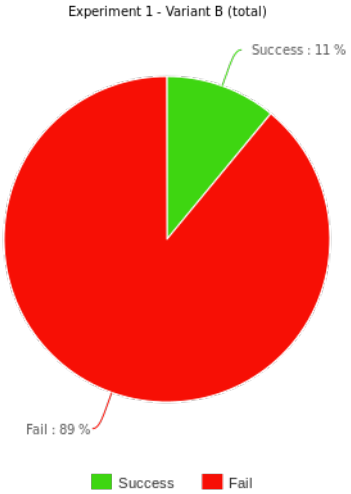

Figure 6.2: Experiment 1, Variant B Total (in %)

	Not Converted (Fail)	Converted (Success)
Variant A	61	17
Variant B	57	7

In appearance, variant A shows a higher rate of conversions with a comparatively close sample size. However, statistical analysis will show whether there is an actual difference between variants. Having plotted the figures obtained from the experiment, a Chi-squared test was performed. The null hypothesis presented was that there is a significant difference between the two groups. The test was performed using an online tool [45]. The resulting one-tailed p-value of this test is $p = 0.0429$ and $\chi^2 = 2.951$, which is lower than 0.05. Therefore, we can reject the null hypothesis, and it can be considered that there is a significant difference among control and treatment variances, with Variant A obtaining a better conversion rate.

6.2.3 Experiment 2

This experiment, in contrast with the first one, and as explained in 6.1.3, was created by the authors. A target audience of mostly young people, aged 18-35, was selected for this experiment, with a total of 612 participants. While most of those participants are based in Europe, some of them are based in North and South America as well as Africa. Also some of those based in Europe have origins in other regions, which adds for a more diverse sample. Out of those 612 subjects, a sum of 166 clicked on the link provided in order to start the experiment. The tracking tool included in the Decision Assistant showed that 91 and 75 participants were respectively redirected to Variant A and Variant B. The conversion goal set for this experiment was to achieve as many successful finishes as possible. The flow created was configured in the Decision Assistant using weights to measure the answers provided by the subjects of the experiment. Upon completion of the flow, and based on the calculations of the final score with the weights of the answers selected, the subject was tested either as successful or failed. Regarding the 91 participants that conducted Variant A, 79%

of them finished the flow and 21% dropped out during the process. The Decision Assistant showed that 65% of participants tested successful , and 14% tested failed. These figures express percentages over the total population. Combining pure fails and dropouts, a total of 35% of participants are hence tested as fail.

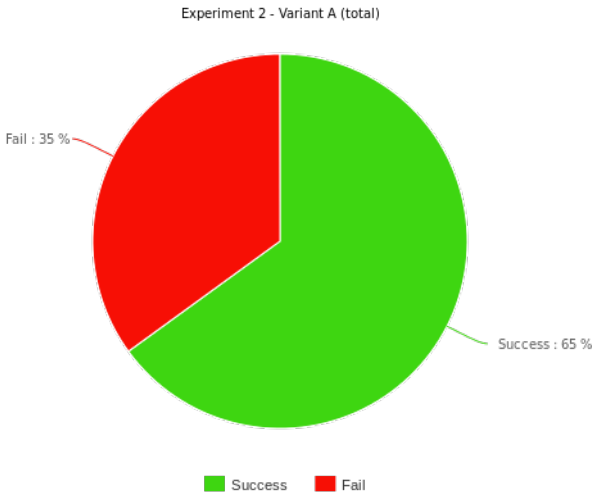

Figure 6.3: Experiment 2, Variant A Total (in %)

Extrapolating the figures to focus only on the results for finished flows, a total of 82% of those were successful, while only 18% were tested as failed.
Regarding Variant B, 75% of the 75 subjects who participated finished the flow, while 25% of them dropped-out. Over the total figures, a 68% of the sample tested as successful, while a comparatively low 7% tested as pure failed. The combined pure fails and drop-outs make the total figure of failures rise to a 32%.

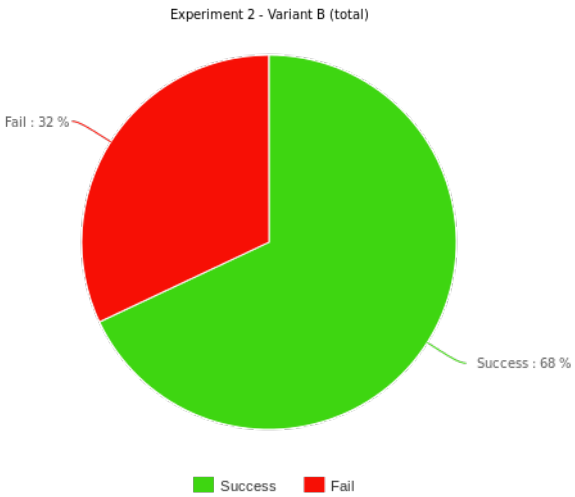

Figure 6.4: Experiment 2, Variant B Total (in %)

As per only-finished figures, a quite high 90% of the participants who finished tested positive, leaving a 10% as failed.

The authors found interesting the high response for both variants. Variant A was presented with a formal language, and it boasted a lower rate of drop-outs. However, even though Variant B presented a more informal language, the fact that the target was predominantly a young audience helped to keep both comparatively high finish and conversion rates. Another characteristic of Variant B was to ask for more private questions such as personal information on economic stability, with the idea of getting an insight on whether this would make participants wary or suspicious about giving this kind of information. Of those who finished Variant A with successful results, eight participants refused to give their personal information, while for Variant B with positive results, all of them provided the personal information requested. One of the participants however, reported that asking personal information was a cause for his drop-out. More discussion about this can be found in chapter 8.
The corresponding contingency table for this experiment can be found below:

	Not Converted (Fail)	Converted (Success)
Variant A	32	59
Variant B	24	51

The test was performed using an online tool [45]. The resulting figures show a sufficiently large sample size and cell count, resulting in a $\chi^2 = 0.184$ and a p-value of $p = 0.3339$ with one degree of freedom. Thus the null hypothesis is accepted and the result for this test is that there is no significant difference between the two groups.

6.2.4 Experiment 3

Out of the 83 contacts that received the send-out, 45 opened the email that contained a link for the Decision Assistant flow. This data was obtained via an email tracking tool provided by the company. The Decision Assistant's internal tracking shows that 15 contacts started the experiment. For this particular experiment, the conversion goal is stated as a contact finishing the flow. Only three contacts finished the experiment, two from Variant A and one from Variant B. It was seen that more people finished the flow when they were not aware of the whole number of steps ahead of them, as presented in Variant A. However, the authors believe it is safe to say that the poor response to this experiment cannot add any significance to this research. More about the possible reasons for the outcome of this experiment can be found in chapter 8 as part of the discussion.
The corresponding contingency table for this experiment can be found below:

	Not Converted (Fail)	Converted (Success)
Variant A	5	2
Variant B	7	1

7
Discussion

This section reports on the answer of the Research Questions of this thesis, based on the results of the Case Study and Experiments conducted in Chapters 5 and 6. Section 7.1 presents a correlation between the results of the case study and the experiments with the research questions that conform this thesis. In addition, Section 7.2 presents a brief discussion along with the limitations found when analyzing the experiments.

7.1 Answer To Research Questions

In this section the authors will present again the research questions proposed in section 3.2. Then, a correlation between the answers from the interviews and the results of the experiments will be performed in order to formulate a suitable answer for the research questions.

7.1.1 Research Question 1

"How can the use of A/B testing be extended from visual aspects of online services in order to optimize sales processes in the E-commerce domain?"

The authors found some evidence from the case study that was performed that can help identify how can A/B testing be extended towards the intended domain. The results of the interviews showed a promising perspective into integrating this type of split testing in the E-commerce domain. All sources agreed on the suitability of the Decision Assistant tool in order to create online sales processes, and they all provided a good insight on what they believe a successful online sales process must offer to the end customer. Among them, the most cited are to provide a believable benefit, an easy to use system, a relation with the customer based on trust in the form of being transparent about your process, and having a good strategy.
The first experiment threw a significant difference between the two tested variants. Variant A seemed to obtain a better conversion rate, something which was corroborated by the statistical analysis of the data gathered. This gives the authors the idea that shorter steps seem to encourage people to engage in completing the process in a successful way. Experiment number two did not show any significant difference among variants, which might lead to the idea that the type of language used to address a specific target group does not play a very important role. However, given the nature of this experiment, the question on whether the authors should check about threats to validity towards the experiment should be addressed. The experiment

was performed anonymously since the participants did not know that they were being subjected to an A/B test, with them thinking that they were only filling out a survey. Even though they were given freedom on when to perform the experiment and whether they should finish it or dropout, the fact that the subjects were perhaps willing to help the authors should be taken into consideration. Another issue might be that the participants, as stated in the description of the experiment, belonged to a relatively young age group, which might make the language variations not so critical.
Based on the evidence found, the authors can say with some reservation that A/B testing can be a valid method in order to test differences in two variants of an online sales process. Out of the three experiments conducted, two were successful in the sense of gathering enough data in order to support conclusions, and the case study provided many insights from the interviewees. The main issues identified is that the processes must be correctly constructed, with a clear goal from the owner's perspective and using the right tools. These are the key factors to be considered in the case of implementing non-visual A/B testing in an online sales process. The failure of the third experiment can be traced back to these issues.
To sum up, A/B testing is a promising instrument for the optimization of sales processes, however more experiments are needed to understand advantages and limitations of using A/B testing in this domain.

7.1.2 Research Question 2

"What could a generalized framework, produced from the aforementioned use of A/B testing, contain so as to be adapted by companies in the field?"

The answer to this question is connected to the partial positive obtained in the first research question. A certain set of characteristics which help optimize sales processes in E-commerce can be identified. Based on the study conducted, A/B testing seems a promising way to test out improvements when conducting online sales processes. However, and given the fact that the answer to the first research question was not entirely positive, the authors believe that instead of establishing a framework, the evidence obtained point more to create a set of guidelines that can be expanded into a future and more complex framework. In order to build this framework, more experimentation should be conducted in order to test more characteristics.
The most cited characteristic to create a successful online sales process was to avoid using irrelevant questions. The irrelevancy of the questions might put users in fear of the real intentions of the owner of the process, such as the intention of acquiring unnecessary data from the users, be it personal data or directly useless information. This request for useless information might also give the user the impression of a poor strategy from the business side, or even worse, the fact that the business is incapable of communicating the features of a product or the details of a service.
From the experimentation it could also be inferred that steps featuring a short number of questions tend to lead to more conversions than hosting a process with fewer, more dense steps. It is worth, nonetheless, to test more extensively this characteristic in order to obtain a better understanding of its benefit in different settings.

Another characteristic that arose from the interviews is the possibility of reordering questions, since it was stated that it is often difficult to come up with a good logical order for them in the beginning. Testing with variants hosting the same questions but organized in different patterns or paths can help to solve this situation. Moreover, the possibility of having different paths for the user adds for more variety in the treatments, which further expands the possibilities of A/B Testing.
Contrary to the expectations, the experiment that had as one of its objectives to test how different language variations (in the way of a more or less formal language) targeted to different groups can play a role, failed to show any significant difference between the two variants. Towards the creation of the future framework, this can be something worth to further test in future experiments by targeting different audiences.
As a conclusion, with the current evidence and data, the authors suggest the aforementioned guidelines into creating successful A/B tests in the studied domain. Chapter 8 will include what kind of future work can be conducted in order to further expand these guidelines, with the objective of constructing a well designed A/B testing framework for E-commerce environments.

7.2 Discussion

The focus of this Master Thesis was to determine whether A/B testing could be extended from the traditional visual aspects. Through the case study and the experiments that were carried out during the writing of this Thesis it was shown that there are positive indications on the suitability of A/B testing experiments that focus on sales processes. The case study conducted with the company personnel concluded that, using a Decision Assistant tool developed specifically for this purpose, A/B testing could be an interesting tool for evaluating sales processes.
The conducted experiments had some positive results as well. The first experiment, conducted in a Business-to-Business setting, showed a significant difference between the two variants, while the second experiment, which was conducted in a Business-to-Consumer setting, didn't have such a significance. A third experiment was also carried out but the responses from its target group were not at a satisfactory level. The first experiment was also successful in getting valuable feedback from participants that either did or did not finish the flow. A big issue highlighted by most of the participants that did not finish the flow was that it didn't interest them, either because they didn't have any time or because the sent-out email seemed like spam. This means that in order to improve the response ratio in a decision assistant flow a lot of consideration must be placed in the format in which it will be presented to potential customers. Also the size of the flow should be as little as possible taking into consideration the hectic lives and the time shortage most people have. Another issue that was noted was that the flow didn't feature an I don't know answer. This reinforces the belief that when making a flow that helps the customer identify his needs, it should be taken into consideration that not all customers are aware of everything that surrounds the product. Providing answers such as "I don't know" could help figuring out the customer's level of knowledge on the subject which in turn can help the decision assistant in providing better results. One participant

cited that he couldn't see the benefit for him doing the flow. This adds to the aforementioned importance of creating a believable benefit for the customer hinted by N1, N5 and N7 in 5.2.3.
The second experiment might pose a risk to whether the results are representative and valid. The sent-out of this experiment was well prepared, however, there is an unusually large proportion of the sample size that finished the flow. This could be interpreted as a willingness from the subjects of the experiment in putting effort into helping the authors carrying out the flow. From the data gathered from the feedback to the experiments a lot of the participants confessed on starting the flow only because of their willingness to help the authors, though most of them admitted to getting interested in the process. Some commented on the language suitability, citing that the language of the decision assistant was not that professional, which indicates that some people could be offended and decide to quit on doing the decision assistant if the language used is not suitable for the target audience. Another piece of feedback was the reluctance of some to reveal personal information which leads to the conclusion that personal information should be gathered only when there is a serious need for them.
The problem with the third experiment could be traced to the fact that it was not correctly thought as an experiment. It is the authors' belief that the flow, although helpful for the company, it was not actually structured in a way that captures the focus of the target group. Most people probably considered it as spam. From the few answers that even the feedback survey got the main problem was that it was not as probable as it should have been. In the end, the authors believe that it was conceived more as a survey for feedback, which means that the person to reply to this would have to be really invested in the product/service to give some of his time and go through the flow.

8

Conclusion

Concluding this Thesis, this chapter begins with Section 8.1 summarizing its key points, finishing with Section 8.2 reporting about future work proposed.

8.1 Summary

This thesis focused on extending the A/B testing, a type of online controlled experiment, from a tool that is used to mosty test the visual aspects of a website to a broader tool that can be used in E-commerce to test the impact of differences in sales processes. For the purpose of this thesis the authors created the Decision Assistant tool, explained in Chapter 3, which implemented native A/B testing features designed specifically for creating variations of a sales process. Through a set of interview the authors gathered insights on what a successful online sales process might be and in which areas could the A/B testing be focused. From there hypotheses where generated and tested by creating three online A/B test experiments. The results of those experiments showed that extending A/B testing in this aspect could be viable, even though some further investigation might be needed. Also the study generated a set of guidelines to follow for the sucessful implementation of such a feat by companies in the field. Those guidelines along with more that could be generated from further experimentation can lead to a generalized framework on using A/B testing as a tool for testing sales processes.

8.2 Future Work

As it was stated in the analysis of the experiments, the more plausible future work that can be conducted with the objective of expanding this thesis work is to extend experimentation in order to try to understand better whether all the proposed characteristics and thoughts collected from the interviews can be applied in expanding A/B testing into E-commerce environments.
The authors suggest experimentation to be carried into a wide variety of target groups, including B2B (Business to Business) and B2C (Business to Consumer) environments. The validation and reinforcement of the suitability of the characteristics for developing optimal sales process through further experimentation would then lead to the construction of a framework that would describe A/B testing in E-commerce environments, something from which companies in this sector could benefit.

Bibliography

[1] G. Linden, "early amazon: Shopping cart recommendations" (2006).
URL http://glinden.blogspot.se/2006/04/early-amazon-shopping-cart.html

[2] R. Kohavi, R. Longbotham, D. Sommerfield, R. Henne, Controlled experiments on the web: survey and practical guide, Data Mining and Knowledge Discovery 18 (1) (2009) 140–181.
URL http://dx.doi.org/10.1007/s10618-008-0114-1

[3] R. Kohavi, R. M. Henne, D. Sommerfield, Practical guide to controlled experiments on the web: Listen to your customers not to the hippo, in: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '07, ACM, New York, NY, USA, 2007, pp. 959–967.
URL http://doi.acm.org/10.1145/1281192.1281295

[4] J. Bosch, Building products as innovation experiment systems, in: M. Cusumano, B. Iyer, N. Venkatraman (Eds.), Software Business, Vol. 114 of Lecture Notes in Business Information Processing, Springer Berlin Heidelberg, 2012, pp. 27–39.
URL http://dx.doi.org/10.1007/978-3-642-30746-1_3

[5] D. Monty, The sales process, in: Sales Hunting, Apress, 2014, pp. 33–38.
URL http://dx.doi.org/10.1007/978-1-4302-6769-0_4

[6] M. McDonald, B. Rogers, D. Woodburn, Key customers: how to manage them profitably, Chartered Institute of Marketing / Butterworth-Heinemann Marketing series, Butterworth-Heinemann, Oxford, 2000.
URL http://eprints.port.ac.uk/7113/

[7] S. L. Vargo, R. F. Lusch, Evolving to a new dominant logic for marketing, Journal of Marketing 68 (1) (2004) 1–17.

[8] K. Storbacka, L. Ryals, I. A. Davies, S. Nenonen, The changing role of sales: viewing sales as a strategic, cross?functional process, European Journal of Marketing 43 (7/8) (2009) 890–906.

[9] C.-A. Chen, J.-C. Jiang, Am I an Online Shopper After All? A Paradox on Definition of e-Commerce from the Consumers' Perspective, Contemporary Management Research 5 (3) (2009) 301–314.
URL http://www.cmr-journal.org/article/viewArticle/1165

[10] J. Andzulis, N. G. Panagopoulos, A. Rapp, A Review of Social Media and Implications for the Sales Process, Journal of Personal Selling & Sales Management 32 (3) (2012) 305–316.
URL http://dx.doi.org/10.2753/pss0885-3134320302

[11] R. C. Erffmeyer, D. A. Johnson, An exploratory study of sales force automation practices: Expectations and realities, Journal of Personal Selling & Sales Management 21 (2) (2001) 167–175.

[12] P. Parvinen, O. Tiainen, J. Salo, E. Pöyry, H. Blakaj, Conceptualizing e-selling, in: Proceedings of the 13th International Conference on Electronic Commerce, ICEC '11, ACM, New York, NY, USA, 2012, pp. 5:1–5:16.
URL http://doi.acm.org/10.1145/2378104.2378109

[13] U. Gretzel, D. R. Fesenmaier, Persuasion in recommender systems, International Journal of Electronic Commerce 11 (2) (2006) pp. 81–100.
URL http://www.jstor.org/stable/27751213

[14] H. J. Ahn, A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem, Inf. Sci. 178 (1) (2008) 37–51.
URL http://dx.doi.org/10.1016/j.ins.2007.07.024

[15] R. Nikolaeva, S. Sriram, The moderating role of consumer and product characteristics on the value of customized on-line recommendations, Int. J. Electron. Commerce 11 (2) (2006) 101–123.
URL http://dx.doi.org/10.2753/JEC1086-4415110205

[16] J. R. Hauser, G. L. Urban, G. Liberali, M. Braun, Website morphing, Marketing Science 28 (2) (2009) 202–223.

[17] O. E. Williamson, O. E. Williamson, The Mechanisms of Governance.
URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.317.5750

[18] W. Wang, I. Benbasat, Trust in and adoption of online recommendation agents, Journal of the Association for Information Systems 6 (3).
URL http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1454&context=jais

[19] S. X. Komiak, W. Wang, I. Benbasat, Trust building in virtual salespersons versus in human salespersons, E-Service Journal 3 (3) (2005) 49–63.
URL http://muse.jhu.edu/journals/esj/summary/v003/3.3komiak.html

[20] S. Komiak, I. Benbasat, Understanding Customer Trust in Agent-Mediated Electronic Commerce, Web-Mediated Electronic Commerce, and Traditional Commerce 5 (1-2) (2004) 181–207.
URL http://dx.doi.org/10.1023/b:item.0000008081.55563.d4

[21] K. Kakavelakis, A. Felstead, A. Fuller, N. Jewson, L. Unwin, Making a sales advisor: the limits of training 'instrumental empathy', Journal of Vocational

Education & Training 60 (3) (2008) 209–221.
URL http://dx.doi.org/10.1080/13636820802305546

[22] R. T. Grenci, P. A. Todd, Solutions-driven marketing, Commun. ACM 45 (3) (2002) 64–71.
URL http://doi.acm.org/10.1145/504729.504730

[23] R. M. O'Keefe, T. McEachern, Web-based customer decision support systems, Commun. ACM 41 (3) (1998) 71–78.
URL http://doi.acm.org/10.1145/272287.272300

[24] C. Huffman, B. E. Kahn, Variety for sale: Mass customization or mass confusion?, Journal of Retailing 74 (4) (1998) 491 – 513.
URL http://www.sciencedirect.com/science/article/pii/S0022435999801055

[25] E. C. Garbarino, J. A. Edell, Cognitive effort, affect, and choice, Journal of Consumer Research 24 (2) (1997) pp. 147–158.
URL http://www.jstor.org/stable/10.1086/209500

[26] Z. Yang, Y. Li, J. Cai, E. Nyberg, Quads: Question answering for decision support, in: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '14, ACM, New York, NY, USA, 2014, pp. 375–384.
URL http://doi.acm.org/10.1145/2600428.2609606

[27] C.-C. Yu, A web-based consumer-oriented intelligent decision support system for personalized e-services, in: Proceedings of the 6th International Conference on Electronic Commerce, ICEC '04, ACM, New York, NY, USA, 2004, pp. 429–437.
URL http://doi.acm.org/10.1145/1052220.1052274

[28] F. Al-Qaed, A. Sutcliffe, Adaptive decision support system (adss) for b2c e-commerce, in: Proceedings of the 8th International Conference on Electronic Commerce: The New e-Commerce: Innovations for Conquering Current Barriers, Obstacles and Limitations to Conducting Successful Business on the Internet, ICEC '06, ACM, New York, NY, USA, 2006, pp. 492–503.
URL http://doi.acm.org/10.1145/1151454.1151528

[29] T. H. Davenport, How to Design Smart Business Experiments, Harvard Business Review 87 (2) (2009) 68–76.

[30] S. W. H. Young, Improving Library User Experience with A/B Testing: Principles and Process, Weave: Journal of Library User Experience 1 (1).
URL http://dx.doi.org/10.3998/weave.12535642.0001.101

[31] F. Fotrousi, K. Izadyan, S. Fricker, Analytics for product planning: In-depth interview study with saas product managers, in: Cloud Computing (CLOUD), 2013 IEEE Sixth International Conference on, 2013, pp. 871–879.

[32] R. Kohavi, R. Longbotham, Unexpected results in online controlled experiments, SIGKDD Explor. Newsl. 12 (2) (2011) 31–35.
URL http://doi.acm.org/10.1145/1964897.1964905

[33] R. Kohavi, A. Deng, B. Frasca, R. Longbotham, T. Walker, Y. Xu, Trustworthy online controlled experiments: Five puzzling outcomes explained, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, ACM, New York, NY, USA, 2012, pp. 786–794.
URL http://doi.acm.org/10.1145/2339530.2339653

[34] P. Hynninen, M. Kauppinen, A/b testing: A promising tool for customer value evaluation, in: Requirements Engineering and Testing (RET), 2014 IEEE 1st International Workshop on, 2014, pp. 16–17.

[35] E. Kaufmann, O. Cappé, A. Garivier, On the complexity of a/b testing, in: In proceedings of the Conference On Learning Theory, Jun. 2014.

[36] Copyright, in: J. Sauro, J. R. Lewis (Eds.), Quantifying the User Experience, Morgan Kaufmann, Boston, 2012, pp. iv –.
URL http://www.sciencedirect.com/science/article/pii/B978012384968700014X

[37] L. Finger, S. Dutta, Ask, Measure, Learn: Using Social Media Analytics to Understand and Influence Customer Behavior, O'Reilly Media, 2014.
URL https://books.google.se/books?id=vZenAgAAQBAJ

[38] E. B. Seufert, Freemium Economics: Leveraging Analytics and User segmentation to Drive Revenue, Morgan Kaufmann, Boston, 2014.

[39] O. Clark, Games As A Service: How Free to Play Design Can Make Better Games, Taylor & Francis, 2014.
URL https://books.google.se/books?id=IWPMAgAAQBAJ

[40] S. Borodovsky, S. Rosset, A/b testing at sweetim: The importance of proper statistical analysis, in: Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on, 2011, pp. 733–740.

[41] A. K. Shenton, S. Hayter, Strategies for gaining access to organisations and informants in qualitative studies., Education for Information 22 (3-4) (2004) 223–231.
URL http://dblp.uni-trier.de/db/journals/efi/efi22.html#ShentonH04

[42] P. Runeson, M. Höst, Guidelines for conducting and reporting case study research in software engineering, Empirical Software Engineering 14 (2) (2009) 131–164.
URL http://dx.doi.org/10.1007/s10664-008-9102-8

[43] R. Yin, Case Study Research: Design and Methods, Applied Social Research Methods, SAGE Publications, 2009.
URL https://books.google.se/books?id=FzawIAdilHkC

[44] C. B. Seaman, Qualitative methods in empirical studies of software engineering, IEEE Trans. Softw. Eng. 25 (4) (1999) 557–572.
URL http://dx.doi.org/10.1109/32.799955

[45] Graphpad software.
URL http://graphpad.com/quickcalcs/contingency1/

A

Appendix I

A.1 The Interview (taken at the company during the first week of June)

- The interview will be conducted in English.
- The interview will be anonymous. No name or reference to any of the participants will be disclosed.
- The interview will be recorded upon permission of the interviewee (it would be extremely appreciated). Transcripts of this recordings will not be published.
- The intended duration of the interview will be at most 30 minutes. However, duration can be flexible on the interviewee's discretion.
- The interview will be semi-structured. This will allow for extra discussion if necessary.

Block 1: Introductions

1. Could you please introduce yourself, including describing your background and position in the company?
2. Did you hear about the Decision Assistant before having the general meeting of the company last Wednesday 27th of May? Are you acquainted with it?

Block 2: Sales processes and DA

3. What, in your opinion, is the key factor for a successful online sales process?
4. How do you think the DA is going to help you achieving that?
5. Among all the customization possibilities that the DA includes (Number and order of steps, number and order of questions, visible progress bar, visible location in the process, visibility of the length of the process, weighted questions allowing more than one possible tracks inside the flow), which ones do you find find more useful to construct a DA flow that in can in turn lead to a successful sales process?
6. How would you combine those customization possibilities in order to produce two variants? (Give example to clarify). Note that this is basically the selection of the characteristics for the flow that will be built and tested.

Block 3: Additional information:

7. Can you tell us a bit more about your target group?

B

Appendix II

B.1 Feedback survey (Sent online to participants of the experiments)

In order to improve the Decision Assistant, we have a few questions for you. It will not take more than a couple of minutes to complete:

1. Did you finish the Decision Assistant Flow? (Yes/No)
2. If no, what made you not finish it?
3. Do you have any suggestions about the tool or the Decision Assistant Flow that you received?