Universiteit Leiden
The Netherlands

Master Computer Science

Stimulating the Adoption of A/B Testing

in a Large-Scale Agile Environment

Name:	Juliëtte Meeuwsen
Student ID:	s1508482
Date:	22/05/2019
Specialisation:	Science-Based Business
1st supervisor:	Dr. M. van Leeuwen
2nd supervisor:	Dr. X. Li

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

In today's business world, software companies like banks focus on making informed decisions in order to satisfy customer needs whilst also succeeding in business. One way to make informed decisions is to adopt A/B testing across the organisation. However, the adoption rate of A/B testing at case company ING—a large international bank— is low on average. In this thesis, we use an exploratory mixed case study to find out how to stimulate the adoption of A/B testing throughout the company. We do this through a survey with 295 respondents, a focus group with 13 participants, and by building a model for investigating the effects of using pre-experiment data. Our findings show that advanced functionalities and high discoverability are key incentives for increasing the adoption rate of the experiment platform at our case company. Last, we provide practical guidelines for improving the rate of significant A/B tests. We show that by using pre-experiment data — user behaviour data prior to the start of an experiment — the power of A/B tests increases.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In most industry sectors digitisation led to a change in operations and management. Previously, companies were focused on identifying and solving technical problems. Now, the focus lies on identifying and solving problems that are relevant to customers and deliver benefits [34]. The change in focus asked for a corresponding change in approach to software engineering; the Lean Startup methodology was defined. The core component of this methodology is the build-measure-learn loop [34]. The first step is figuring out the problem that needs to be solved and then develop a minimum viable product (MVP) [28]. Once this MVP is created, it can be used in an experiment to collect measurable data. The experiment results are used to validate or invalidate the initial solution and then one makes a decision on whether to move forward to the next stage, test the product on a new problem, or stop [30].

1.1 A/B Testing

One form of experimentation is A/B testing (or randomized experiments, controlled experimenting, control/treatment tests) [25]. Randomised trials are now a standard procedure in medicine development [55]. Although A/B testing has been patented only this century [15], historical events show that it is not that new. James Lind, a Scottish surgeon, found a remedy for scurvy — a disease which often struck sailors on long voyages — by testing the effects of providing different types of fruit to a group of patients in a clinical trial. He found that those patients that ate citrus fruits recovered much faster than those who where given other types of food [35].

A/B testing is a process to isolate and test aspects of a product or service that impacts its effectiveness [8]. As such, it allows defining a causal relationship with high probability. The applications are endless, and nowadays it is often used for marketing and user experience improvements. The most basic setup of an A/B test is to evaluate one factor between two variants, respectively a control (version A) and a treatment (version B) [29]. The control is usually the default version and the treatment is the change that is being tested. Figure 1.1 shows the high-level structure of an A/B test. Splitting all traffic in half (50% of the users get the control variant, whereas the other 50% sees the treatment) provides the experiment with maximal statistical power. The analysis of such experiments assesses whether the statistical distribution of the treatment is different from that of the control level. Before executing an A/B test,

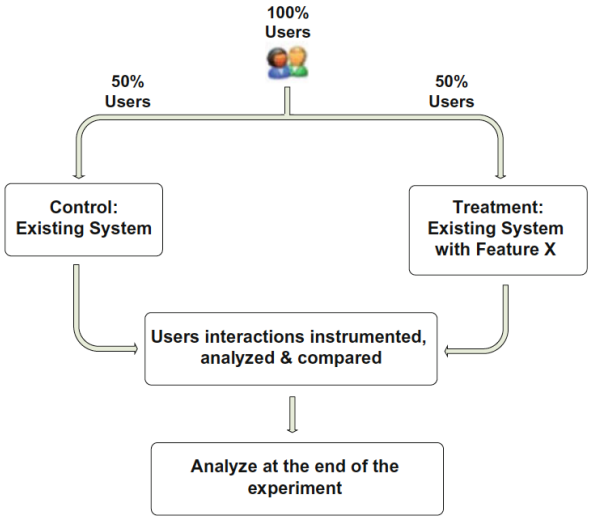


Figure 1.1: High-level structure of an A/B test (taken from [8]).

various aspects of planning are required: estimating adequate sample size, gathering the right metrics, tracking the right users, and randomisation unit. Analysing the experiment results is a straightforward application of statistical methods, such as hypothesis-testing procedures, re-sampling/permutation methods, and diagnostic tools.

1.2 Problem Statement

Software development companies recognise A/B testing as an effective method to learn what customers really value, however they seldom succeed in evolving and adopting the methodology [8] [16].

This study is performed at ING, a large FinTech company that strives to be data-driven. Yet, the adoption rate of the internal experiment platform designed specifically for A/B testing is not as high as desired; only 28 out of a potential of 373 teams use it. Moreover, only 15.1% of all experiments executed using the experiment platform have a significant test outcome (using the Chi-Squared test [37] with a confidence level of 95%). The aim of this thesis is to perform a mixed method study with both quantitative and qualitative components on how an A/B testing platform in a large-scale agile environment can operationalise large-scale experimentation. This approach leads to the following research questions.

RQ 1: What factors influence the usage of continuous experimentation in cross-functional teams?

Tracking down reasons for usage of continuous experimentation and the experiment platform could be key for either further spreading the usage of continuous experimentation or defining improvements of the experiment platform itself. Our expectation is that composition of the team and experience with experimentation culture influence the usage of continuous experimentation. Other reasons could be that the experiment platform does not add value to teams (e.g. not time reducing or too much effort needed), or cross-functional teams do not even know that it exists. From this the following sub-questions are defined:

RQ 1.1: *What are motives for teams to use the experiment platform?*

RQ 1.2: *Why do teams not use the experiment platform?*

RQ 1.3: *What actions can be taken to increase the adoption of the experiment platform in cross-functional teams?*

RQ 2: What can be done to increase the power and sensitivity of A/B test outcomes?

Once it is verified that the experiment platform consists of powerful and sensitive algorithms such that experiments actually have a purpose and can add value to a team, we expect it to be easier to convince teams to use the experiment platform. This means that if the amount of significant experiments executed in the experimentation platform were higher than in its current stage, it would be easier to convince new users of the effect of the platform. We decide to test the effect of using pre-experiment data — user behaviour data prior to the start of an experiment — on the power of A/B tests. This research question is refined as follows:

RQ 2.1: *To what extent does using pre-experiment data for A/B tests affect test outcomes?*

RQ 2.2: *What other methods exist that could potentially improve the experiment platform?*

Finding answers to these questions will help to better understand the adoption of continuous experimentation in a large-scale agile environment, and more importantly, the adaption of a specific in-house infrastructure across cross-functional teams. According to a survey of more than $10\,000$ business across 140 countries, 94% of the respondents report that agility and collaboration are critical to their organizations success [13]. The findings will be useful for these large-scale agile environments and provide insights on how to operationalise large-scale experimentation using an internal platform for A/B testing. In summary, this thesis makes the following main contributions:

- We investigate factors that influence the usage of and motives for continuous experimentation in a large-scale agile environment through a focus group and survey.
- We investigate the effect of using pre-experiment data to improve power of A/B tests using simulation of our proposed model. The implementation of the model can be found at `https://github.com/Hearrtbeatt/model-simulation`.

1.3 Thesis Overview

This Master thesis project is supervised by Dr. M. van Leeuwen and Dr. X. Li and is executed at ING Netherlands and The Leiden Institute of Advanced Computer Science (LIACS).

In Chapter 2 related work is mentioned. The experiment platform and approach for analysis will be explained in Chapter 3. Obtained results are presented and evaluated in Chapter 4. Finally, in Chapter 5 conclusions will be drawn about the project and suggestions for future work will be mentioned.

Chapter 2

Related Work

For more and more companies experimentation is key to innovation and growth. At Google [14], for example, every change that potentially affects its user experience is tested. Google's design goals in the experiment infrastructure are *more, better and faster*. More experiments need to run simultaneously, bad experiments need to be identified and stopped quickly, and it should be easy and quick to set up an experiment, even for a non-engineer without any coding experience. An overlapping infrastructure is used for scaling more experiments faster and simultaneously. Tools are used for data file checks, real-time monitoring and fast and accurate analysis. Tools and infrastructure are important for implementing experimentation successfully, however they find education also essential for facilitating robust experimentation.

Booking.com [39] is a company where online controlled experiments have been used for more than ten years to conduct evidence-based product development. The infrastructure shows resemblance with infrastructures that Microsoft, LinkedIn, and Facebook use for experimentation, however there are also differences supporting Booking.com's aim for decentralising experimentation at scale; a repository of past failures and successes enables anyone to form new hypotheses, accurate customer behaviour measurements build trust in experimentation, and experiments are made accessible and safe by designing extensible frameworks with built-in safeguards.

According to Fabijan et al. [8], software development companies are aware of the competitive advantage continuous experimentation provides, however they seldom succeed in evolving and adopting the methodology. Therefore, a case study in a large software-intense company (Microsoft) with a highly developed experimentation culture resulted in a complete model for developing and scaling continuous experimentation; the Experimentation Evolution Model. The Experimentation Evolution Model helps to move from a situation with ad hoc data analysis towards continuous controlled experimentation at scale. It lists a number of prerequisites for experimentation as well as three dimensions of evolution. The technical evolution dimension focuses on the technical part of the experiments, such as complexity, pervasiveness of development activities, and overall focus of development activities. The organisational evolution dimension is all about the organisation of the data science teams and their self-sufficiency for experimentation. The business evolution dimension deals with the overall evaluation criteria.

Small differences in key metrics, on the order of fractions of a percent, can have significant business implications, as Deng et al. [4] found out at Microsoft. CUPED is an approach that utilises data from the pre-experiment period to reduce metric variability and hence achieve better sensitivity. Unfortunately this approach is not always applicable, for instance if the purpose is to measure retention rate or test on new users. Netflix [5] also recommends using a post-assignment variance reduction technique like CUPED over at-assignment variance reduction techniques like stratified sampling in large-scale controlled experiments. Reason for this is that stratified sampling performs worse than post-assignment techniques. Moreover, post-assignment techniques are cheaper to implement and very flexible.

Although the largest software companies such as Google, Microsoft, Facebook, Netflix do publish about A/B testing in their environment, most relevant information can be found on informal and non peer-reviewed channels, such as blogs on so-called research web pages [7], [22], [27].

According to Ahmed et al. [2], internal marketing helps the process of identifying current behaviours and probes why they are occurring. Once specific employee behaviour patterns have been established, it is then possible to create specific internal marketing programmes to induce behaviours for enhanced implementation. Involvement and commitment combined with clear sense of purpose are pre-requisite for the of adoption of new products and/or services. The most common application of internal marketing is creating communication strategies. Each employee behaves differently, and can be assigned to a specific employee behaviour pattern. Each of these different types of employees have different needs, and all these needs have to be taken into account when it comes to internal marketing.

Chapter 3

Methods

In order to better understand which factors affect the usage of API-based A/B testing in a large-scale agile environment and the effects of using other methodologies for algorithms in the experiment platform an exploratory mixed case study design is used. Data of the experiment platform is analysed, results are substantiated using a focus group and a survey. A model for using pre-experiment data to improve power of A/B tests is designed, applied and simulated.

3.1 Case Study Research

A case study is an effective methodology for investigating and understanding complex issues in real world settings [31]. It investigates contemporary phenomena in its natural context, which makes it well suited for the aim of this research [44]. This case study has been performed at ING [21]: a large Netherlands-based bank with 36 million customers in 40 countries and 51 000 employees, of which 15 000 working in IT. The bank is currently in the middle of a technology shift from a pure finance-oriented to an engineering-driven company. To accelerate innovation ING uses its own methodology, which combines Lean Start-up, Agile Scrum, and Design Thinking methods, and encourages fast experimentation based on customer feedback. For the latter, ING has built its own experiment platform for A/B testing.

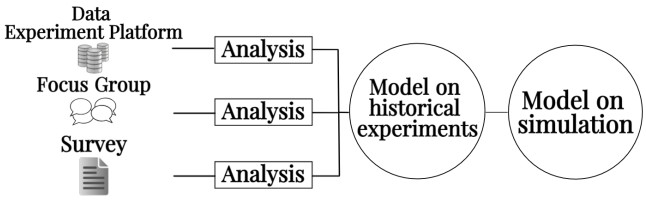


Figure 3.1: Abstract illustration of research design.

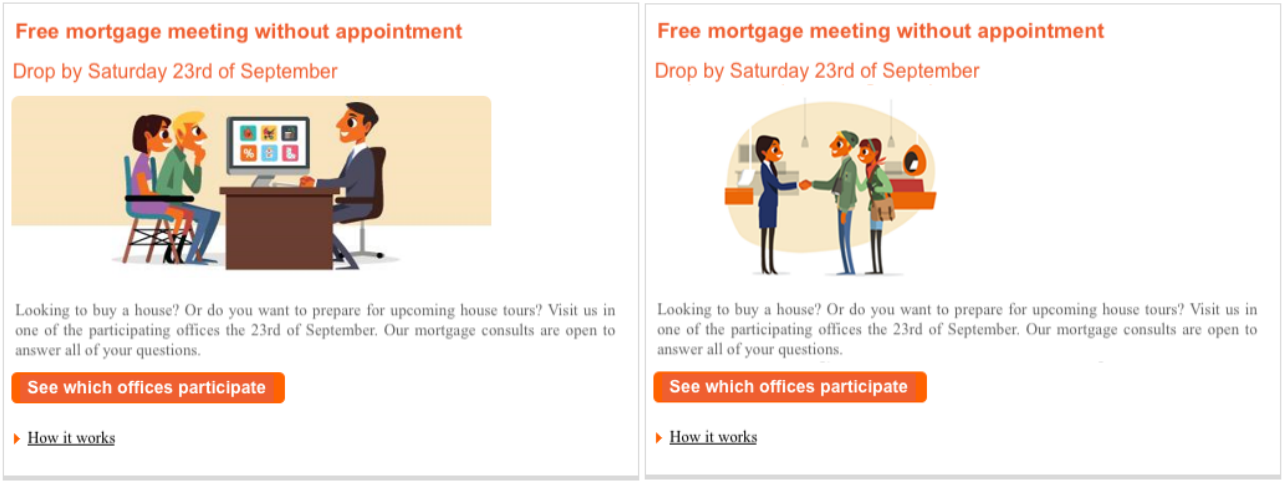
The experiment platform stores all experiments in a dashboard which can be used for data analysis. The implementation and historical data are used for research on the power of the experiment platform's algorithms. Additional perceptions in motives, key principles and usage are provided by a focus group and a survey. See Figure 3.1 for an illustration of the research design.

3.2 Experiment Platform

ING decided to build its own solution for API-based A/B testing, to be able to experiment with and easily connect to other systems in the ING environment. Using an internal tool has several advantages such as customisation, independency and efficiency [23]. The experiment platform connects to commercial tools for content management and digital analytics, complemented with a self-built API solution for connecting ING's background systems. The solution can be used for running A/B tests on different channels within ING, for instance the website.

A typical example of an A/B test on the website is the Visual Mortgages-day test (see Figure 3.2). This example shows two variants of a page inviting visitors to Visual Mortgages-day. The purpose of this A/B test was to optimise the *Next Best Action* (NBA) — usually a personal offer — on the ING website regarding mortgages. The hypothesis for this test was *'If an NBA for a mortgages-day contains a deviating (round) formatted image, this leads to more clicks, because the expression is more striking and the form is a visual cue towards the text and the button'*. When the hypothesis is confirmed, the alternative — Figure 3b — is used during the remaining campaign. If the hypothesis is rejected, then the default — Figure 3a — is used, and further testing is done with alternative texts or other selections. The outcome of this test specifically was that a round visual with people standing (Figure 3b) performed 28% better than the rectangular visual with people sitting at a desk (Figure 3a), with p-value 0.0438, and confidence level 95%.

Figure 3.2: An overview of the Visual Mortgages-day test.



(a) Rectangular visual, sitting

(b) Round visual, standing

Creating a new experiment using the experiment platform is fairly straightforward; first the URLs to different variants are provided, then the purpose is explained by defining the hypothesis and expected results. Lastly, target groups and conversion point are selected. The API itself handles the randomisation and deals with tracking, measuring and representing experiment data. All statistical tests are performed in the background and visualised near real-time in the interface. The experiment platform shows user data of the number of visitors, difference between variants and statistical test results. Via the experiment platform teams are in complete control of managing their own running experiments, although it is advised to run experiments for at least one week because experience shows that results can easily change over time due to time-based decisions (e.g. receiving salary at a specific date). The main advantages of a decentralised approach [39] are integration with own environment and the ability to run experiments without the

assistance of others. Disadvantages are prioritising experiments and education. A decentralised approach is used for the platform, which means that each team needs education to gain confidence in continuous experimenting and using the platform.

3.2.1 Data Collection and Exploration

The experiment platform stores all information about the experiments executed by a member from a cross-functional team, such as time span, hypothesis, outcome(s) and domain. Our study is based on historical data collected at the company starting May 2016 till the end of October 2018. For the collected data, the total amount of experiments and teams is determined. For testing the effects of pre-experiment data only experiments with status "Ended" are considered. The *Click-Through-Rate* is chosen as success rate since "number of clicks" usually is the measuring point. The Click-Through-Rate defines the number of successes against the total. On web pages this usually is measured as the number of clicks on a certain button divided by the total number of visitors. Traffic and "number of clicks" for each page connected to an experiment are retrieved from *Webtrekk Analytics*[1].

3.3 Focus Group

3.3.1 Design

In order to gather information about knowledge and experiences across different departments, as well as identifying potential improvements of the experiment platform, a focus group is used. The main advantage of a focus group is that each individual brings their own view on the situation, which can lead to discussions. These discussions help to explore and retrieve valuable insights [38]. The session consists of half an hour brain-drawing [51] followed by a guided discussion about likes, dislikes, and recommendations for the experiment platform.

3.3.2 Selection of Participants

In order to identify needs for users of the experiment platform, participants are those who use the experiment platform or stated to be interested in using it in the near feature. With the ideal size for a focus group in mind [36], fourteen participants that most likely have different visions to stimulate discussion (e.g. various backgrounds, ages, genders, and teams) are invited.

3.3.3 Focus Group Operation

The session is hosted by the author, accompanied by a colleague internal at ING. The whole session is audio-recorded for further reference. The session starts with a brief introduction of the experiment platform. Then participants write

[1]Webtrekk is a customer intelligence platform that allows to connect, analyse and activate user and marketing data across all devices.

down likes and dislikes about the current system. The session continues with the brain-drawing part, where participants individually capture their vision of an ideal experiment platform on paper. Afterwards the piece of paper is handed over to their neighbour who has to interpret and add to the idea. This round was repeated once. The brain-drawing ends by retrieving the initial idea and take in what others added. The session concludes with a discussion; drawings are discussed and an "ideal" experiment platform is described.

3.3.4 Data Analysis

The recordings of the session are transcribed manually. Comments (likes, dislikes and recommendations) are categorised using RQDA [20]. Collaboratively the final set of categories is determined using categorisation. This set of categories is used to code all comments, including those who had been used to find the categories. Categories are ranked, based on occurrences, to identify the most popular categories.

3.4 Survey

3.4.1 Design

The aim of the survey is to retrieve as much information about a vast amount of teams across ING NL and Belgium, such as experience with, perception of and vision on experimentation. During earlier meetings with colleagues at ING, several reasons for the lack of experimentation where identified (see Figure 3.3). The main goal of the survey is to verify whether or not these reasons are valid across a larger population.

Figure 3.3: Graphical representation of reasoning behind the lack of experimentation at ING.

The scope of the research questions is reflected in the survey questions and consists of either dichotomous, single-choice, multiple-choice, matrix or slider bar questions. To further explore opinions and views of respondents, an optional field for all slider bar questions is included.

Q1. *Is experimentation part of your team's development cycle? (Slider bar: 0 being no part at all, 10 being fully integrated).*

Q2. *Have you heard of the experiment platform before?* (Dichotomous: yes/no)

Q3. *I wish my team would use experimentation more often to measure customer experience.* (Matrix: Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree, Don't Know)

Q4. *At least two members of my team are passionate about experimentation.* (Matrix: Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree, Don't Know)

Q5. *The experiment platform is easy to use.* (Matrix: Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree, Don't Know)

Q6. *How many A/B tests using the experiment platform are executed within your team?* (Multiple-choice: 0, 1, 2, 3, or ≥ 4)

Q7. *What kind of education do you think is needed before you can use the experiment platform?* (Multiple-option: e-learnings, documentation, tutorial, or workshop)

Q8. *How likely is it that you recommend using experimentation to a colleague? (Slider bar: 0 being not likely at all, 10 being very likely).*

Q9. *How likely is it that you recommend using the experiment platform to a colleague? (Slider bar: 0 being not likely at all, 10 being very likely).*

3.4.2 Selection of Participants

In order to explore differences, the survey is sent out to members of teams independent of whether they are using continuous experimentation. To ensure data collection across a large number of diverse projects, members of teams active in 2018 are selected. In total 1461 participants belonging to 373 teams are contacted. The participants are identified based on their team's purpose (e.g. working on front-end, analysing customer behaviour or improving customer experience). We decide to only include those teams for whom using the experiment platform is most relevant, respectively Customer Journey Experts (CJE), Data Analysts (DA) and Engineers (ENG).

3.4.3 Survey Operation

The survey is uploaded into *Collector*, a survey management platform internal to ING. An invitation letter describing the purpose of the survey and how results provide valuable insights in adoption of continuous experimentation is sent to the selected participants. Respondents have a total of 12 days to submit their answers, they are reminded once. The survey ran from 4 January 2019 to 15 January 2019.

3.4.4 Data Analysis

Survey questions have been peer-reviewed. Comments (likes, dislikes and recommendations) are categorised using RQDA [20]. Collaboratively the final set of categories is determined using categorisation. This set of categories is used to code all comments, including those who had been used to find the categories. Categories are ranked, based on occurrences, to identify the most popular categories. The analysis includes a general overview of all responses as well as a separate analysis for each of the three roles in order to compare different visions. For the latter step, we build stepwise logistic regression models to identify whether having a specific role influences the responses for individual questions. For each of the Likert-scale, multiple-choice, and multiple-option questions, we build a model with the presence of a Strongly Agree response (yes/no) as dependent variable and the roles (DA, ENG or CJE) as independent variables [6]. We build similar models for the presence of Agree and Disagree responses. In total, we build 27 models, three for each of the 6 questions. We remove roles for which the coefficient in the logistic regression model was not statistically significant at $p < .01$. In order to find differences for answers on slider bar questions, a box plot in combination with the ShapiroWilk (SW) test is used. The SW-test provides a means of testing whether a set of observations are from some completely specified continuous distribution, and is a powerful normality test [24].

3.5 Statistics

For determining which of the variants performs best, the experiment platform currently uses two statistical calculations. Statistical significance, the likelihood that the difference between two groups could be just an accident of sampling, is usually calculated as a 'p-value' [42]. In addition to this the experiment platform has a built-in conversion change measurement, which determines the effect of the variant relative to the control.

3.5.1 Statistical Inference: Significance Tests

A hypothesis is a statement about a population [3]. It is usually a prediction that a parameter describing some characteristic of a variable takes a particular numerical value or falls in a certain range of values. Significance tests use data to summarise the evidence about a hypothesis by comparing point estimates of parameters to the values predicted by the hypothesis. An alpha level is the probability of a type I error i.e., the error that is made by rejecting the null hypothesis when it is true. Beta is the opposite; the probability of rejecting the alternate hypothesis when it is true (type II error), see Table 3.1.

Table 3.1: Statistical hypothesis testing.

Decision	H_0 **is true**	H_0 **is false**
Retain H_0	Correct conclusion $(= 1 - \alpha)$	Type II error $(= \beta)$
Reject H_0	Type I error $(= \alpha)$	Correct conclusion $(= 1 - \beta)$

Two-Sample Test for Proportions

Let π denote the proportion for the population, let $\hat{\pi}$ denote the sample proportion and, let x_1, x_2 be the number of successes and, let n_1, n_2 be the total number of visitors for respectively sample 1 (control) and sample 2 (treatment). The five parts in a significance test method for two proportions:

1. **Assumptions** — type of data, randomisation, population distribution, sample sizes
2. **Hypotheses** — two hypotheses; null hypothesis and alternative hypothesis. The null hypothesis $H_0 : \pi_1 - \pi_2 = 0$, where π_x denotes a particular proportion value between 0 and 1. The most common alternative hypothesis is $H_a : \pi_1 - \pi_2 \neq 0$. This two-sided alternative states that the population proportions differ from the each other.
3. **Test statistic** — the parameter to which the hypotheses refer has a point estimate, which summarises how far it falls from the parameter value H_0. Thus, the number of standard errors between estimate and H_0 value.

   The sampling distribution of the sample proportions has standard error:

   $$se_0 = \sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{3.1}$$

   with pooled estimate:

   $$\hat{\pi} = \frac{x_1 + x_2}{n_1 + n_2} \tag{3.2}$$

   To compare the two proportions a z-test is used:

   $$z = \frac{\text{difference between sample proportions}}{\text{standard error}} = \frac{(\hat{\pi}_2 - \hat{\pi}_1)}{se_0} \text{ , where } \hat{\pi}_i = \frac{x_i}{n_i} \tag{3.3}$$

4. **P-value** — probability summary of evidence against H_0. The probability that the test statistic equals the observed value or a value even more extreme in the direction predicted by H_a. It is calculated by presuming that H_0 is true. The smaller the p-value the stronger the evidence is against H_0. The p-value is a two-tail probability for $H_a : \pi_1 - \pi_2 \neq 0$.
5. **Conclusion** — Interpret what the p-value tells and make a decision by rejecting or not rejecting the null hypothesis. The smaller the p-value, the more strongly the data contradict H_0 and support H_a. A defined significance level α is the probability of a test rejecting the null hypothesis given that it were true. The p-value of a result is the probability of obtaining a result given the null hypothesis were true. H_0 is rejected (making a test result statistically significant) if $p \leq \alpha$ for a pre-specified α-level. The most often picked value for α is 0.05.

Power Analysis

Power is assigned by 1 - β and therefore, is the probability of rejecting H_0 when it is false. This basically means, given a real effect what is the likelihood that an experiment will yield a significant result (e.g. if the power is 41%, a real effect is missed 59% of the time). Power analysis is important, because it helps planning the study, saves time and money, and helps to reasonably allocate resources [47]. The power level is affected by:

1. Sample size
2. Effect size
3. Significance level (probability of avoiding Type I error)
4. Power (probability of avoiding Type II error)

Given any three, the fourth can be determined using a statistical power analysis. One way to do so is using the `R` `pwr`-package which implements power analysis as outlined by Cohen [41]. Since the calculation for pre-experiment data consists of two proportions with unequal sample sizes, pwr.2p2n.test() is used for calculating power. In order to determine the sample sizes for the normal case, that is equal sample sizes, pwr.2p.test() is used.

Significance Test in Experiment Platform

Currently in the experiment platform, the statistical significance calculation only works for tests with one control and one variant group; multivariant tests cannot be calculated. Pearson's Chi-square Test with a $\alpha = 0.05$ is used for calculating significance. Pearson's Chi-square Test is designed to analyse categorical data and tests how likely it is that an observed distribution is due to chance. Basically, it tests the null hypothesis of whether variables are independent.

The Chi-squared value is calculated by Forumula 3.5 [1]. Where O represents the observed frequency, E the expected frequency where independency is assumed. Observed frequencies are real observations from experimental data. Expected frequencies are calculated using probability theory in a contingency table using Formula 3.4. For each entry in the table the total in the ith row and jth column is divided by the table grand total. For each single data item in the data set the Chi-squared value is calculated, hence the summation from 1 to n. The Chi-squared value serves as input for the p-value. Along with the pre-set degrees of freedom and chosen significance level, a p-value can be determined.

$$E_{ij} = \frac{T_i + T_j}{N} \tag{3.4}$$

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{3.5}$$

Both the Chi-Square Test and the z-test are used for testing independence of variables. Although the Chi-Squared test generalises the two-sample z-test to a situation with more than two proportions, in case of comparing two proportions the Chi-Squared test is equivalent to Z-test for two proportions. Both assume that the sample is to be is normally distributed and there is random sampling. A limitation of the Chi-Square Test is that it is sensitive to either very small or large samples [17]. The Chi-Squared Test is a non-parametric test, whereas the z-test is a parametric test. The advantages of using a non-parametric over a parametric test [49] are that small sample sizes are acceptable, have fewer assumptions and can be used for all data types, including nominal variables. However, non-parametric tests are less powerful than parametric tests if assumptions have not been violated and could require a larger amount of time for completing calculations.

Limitations of Significance Tests

Significance tests do not say anything about the size of the effect, therefore statistical significance is not the same as practical significance. A test may not indicate the plausibility of a particular value in H_0, it is unclear whether and which other potential values are plausible. A solution could be to use a confidence interval, which helps to determine the rejection of H_0 has a practical importance by displaying the entire set of plausible values. Other limitations are only to report results if the tests are statistically significant, tests that are by chance statistically significant and interpreting the p-value as the probability that H_0 is true.

3.5.2 Other Calculations

Conversion Change

This method determines the effect for each all variants relative to the control group. This means which variant has the highest change of conversion. For the experiment platform the conversion change is based on the conversion ratio; number of successful clicks versus the total number of visitors on the responding web pages. **Example:** Let variant A have a conversion ratio of $3/10$, and variant B a conversion ratio of $7/10$. Then, variant B has a 133.33% higher conversion ratio, thus a conversion change of 133.33%.

Standard Deviation

Standard deviation [53] is a measure of dispersion of a set of data values. It is used to indicate the spreading of data around the mean. Hence, a low value for the standard deviation indicates that the data points are close to the mean, and vice versa. **Example:** Let the data define the height of a human being in the Netherlands. On average males are 181.1 m, with a standard deviation of 7.42 m. Females are 169 m, with a standard deviation of 7.11 m. The majority of the population will fall in these areas. For example, a female of 174.2 m, still belongs to the majority, whereas a male of the same height is an outlier.

Confidence Level and Confidence Interval

The confidence level [50] indicates the probability for a statistical parameter that repeating the same test will produce the same results. The confidence interval is based on mean and standard deviation. For example when a test takes into account only a small subset of a large population, the probability describes the likelihood of retrieving the same results when generalising the test. Naturally, the factors that affect confidence intervals are population size, sample size and percentage. **Example:** Let the confidence interval be equal to 95% (this is a range of values that for 95% certainly contains the true mean of the population). So for a test which has a confidence level of 95% and a confidence interval between 3 and 6 means that for 95% of all cases the result is within the range of 3 and 6.

3.6 Using Pre-Experiment Data to Improve Power

Pre-experimentation data can increase the accuracy of test results, that is increasing the sensitivity of online experiments for more precise assessment of values [4]. Hence it has been decided to investigate the effects of pre-experiment data on A/B tests in the experiment platform. We use data of number of clicks in combination with the number of visitors per web page per day for a period of 14 days prior to the experiment start date. Instead of continuous metrics, which is a prerequisite for implementing CUPED (see Section 2), our data consists of discrete variables; in most cases visitors on our web pages, such as in Figure 3.2, do not return frequently. Therefore we propose a new approach that includes pre-experiment data for A/B tests with binary outcomes.

The calculations as described in Section 3.5.1 (Formulas 3.1 and 3.2) are adjusted such that pre-experiment data is taken into account. Let x_3 be the number of successes and n_3 be the total number of visitors in pre-experiment data. We assume that the control/treatment situation is equal to the situation before the experiment run, thus that the conversion rate is stable over time. Moreover, we assume that sampling is random from the population. Now the standard error is calculated as follows:

$$se_0 = \sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1+n_3}+\frac{1}{n_2}\right)} \tag{3.6}$$

with pooled estimate:

$$\hat{\pi} = \frac{x_1+x_2+x_3}{n_1+n_2+n_3} \tag{3.7}$$

The expectation is (with the test statistic being calculated as in Formula 3.3 and $\hat{\pi}_1 = \frac{x_1+x_3}{n_1+n_3}$) that because of the increase in data, the differences between the two proportions is magnified such that the power of the test increases.

3.7 Evaluation of Pre-Experiment Model

An analysis of API-based A/B tests is performed in order to find out whether the experiment platform can detect small ($< 15\%$) differences in conversion ratio. Moreover, a simulation of the model is performed to generalise results.

3.7.1 Analysis of A/B Test Outcomes

Firstly, the implementation of the current A/B tests in the experiment platform is inspected, and specifically the sensitivity of significance threshold. Out of all available experiments in the experiment platform, only those that have been finished are considered. The p-value is calculated using the upper tail probability of the normal distribution, with the z-score as input: st.norm.sf(abs(float(zscore)))$*$2 [11]. A two-tailed test is used, since it leads to more accurate and reliable results than when using a one-tailed test. In case of A/B testing one cannot be sure of the direction of the difference in the key metrics.

Secondly, to verify sensitivity it is tested what the effect is of pre-experiment data on the power and outcomes of A/B tests. Calculating power allows to determine the probability of detecting an effect of a given size with a given level of confidence, under sample size constraints. If the power is unacceptably low, the A/B test results are not reliable enough due to the high probability of false negatives. For power analysis along the lines of Cohen [40], a function especially designed for two proportions with different sample sizes is used, namely pwr.2p2n.test() [43]. This function out of the `R` `pwr`-package takes as input the sample sizes, effect size and chosen significance level to determine the power for comparing two proportions with unequal sample sizes. Cohen suggests that h values of 0.2, 0.5, and 0.8 represent small, medium, and large effect sizes respectively.

3.7.2 Simulation

A simulation of the model is used to identify the characteristics of the data suitable for this model. The factors that are taken into account is the amount of traffic, the effect size and absolute proportions to answer the following questions; what is most suited for this model and which amount of pre-experiment data is most effective? For this, six variables are varied, respectively π_1, π_2, π_3, n_1, n_2 and n_3. x_i can be computed directly from the previously mentioned variables. In order to simulate the variance of a sample, a multinomial distribution is used with π_i as probability and n_i as the number of instances. **Example:** Let $n = 5$ and $\pi = 0.4$. The result of 5 instances (in this case coin flips) could then be $0, 1, 1, 1, 0$, where 1 denotes "head", and 0 "tails". Thus, x is equal to 2, and $\hat{\pi} = \frac{3}{5} = 0.6$.

We assume an equal split of traffic, hence $n_1 = n_2$. Moreover, we assume that the pre-experiment version is equal to control, hence $\pi_3 = \pi_1$. π_1 varies from 0.1 to 0.9 with step size 0.1. The effect size is varied for each value of π_1. π_2 is calculated by $\pi_1 +$ effect size. The absolute effect size is used for determining differences between variants, hence only positive effect sizes are used ($0, 0.001, 0.002, 0.01, 0.02, 0.1$, and 0.2.). Besides the p-value calculation, a power analysis is performed. In order to determine the power for all possible combinations of the settings, again pwr.2p2n.test() is used, with the same input as described in Section 3.7.1. In addition, the function pwr.2p.test() [43] (with as input preferred power, chosen significance level, and effect size) is used to determine the minimal size of the samples needed.

3.8 Threats to Validity

Internal Validity. To mitigate the risk of bias and inaccurate responses, established guidelines for designing and executing the survey were followed [48] [54]. The survey was kept as short as possible and only one reminder was used. The danger of a focus group is group consensus; to prevent this participants were asked to first write down their ideas before discussing with others. A known risk of a coding process is the loss of accuracy of the original response, due to an increased level of categorisation [33]. To mitigate this risk, it was allowed to assign multiple codes to the same answer. The experiment platform has been in use for two years, this might not be enough to say something about adoption of this technology. The majority could be early adopters, but this can only be stated with more insights in the technology life cycle [45] of the platform.

External Validity. The main result of this thesis entails advice for stimulating adoption of an experiment platform internally at ING. This research is conducted in collaboration with employees at the case company. Although this set-up enabled continuous access and insight, the risk of bias means that the contributions of this thesis cannot directly be translated to other companies [44]. To mitigate this risk, simulation of the model has been used such that no company-specific data is taken into account. Nevertheless, replication of this work in other organisations to get more general conclusions is advised.

Chapter 4

Evaluation

This chapter presents the results of the analysis of the data from the experiment platform, the focus group session and the survey, and the results of applying the pre-experiment model to historical data and the simulation of the model.

4.1 Experiment Platform: Growth in Usage

An analysis of the experiment platform data provides insights into the usage of the experiment platform. Since the introduction of the experiment platform the total number of experiments has increased; 15 experiments where executed in 2016, 21 in 2017 and 131 in 2018, with an average of 1.25 experiments per month in 2016, to 1.75 in 2017, and 10.92 in 2018 (see Figure 4.1). An interesting aspect is the number of teams. This is also growing over time, namely, 9 teams used the experiment platform in 2016, 11 teams in 2017 and 14 teams in 2018. In total 28 unique teams have used the experiment platform at least once. What is most remarkable is that out of the 196 entries that exist in the experiment platform only 116 experiments (60% of the entries) have a real purpose; with a well-defined hypothesis, version descriptions and title. Other experiments seem to be focused on trying out the experiment platform rather than actually executing an A/B test.

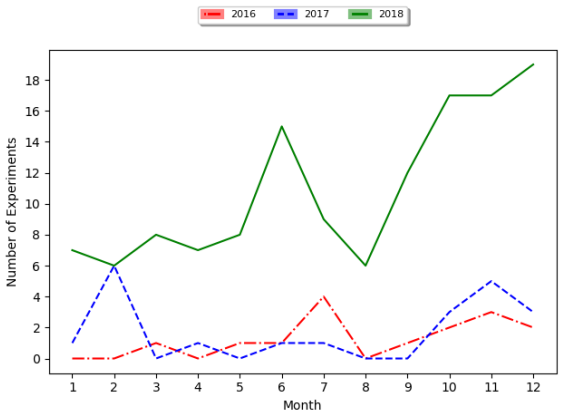


Figure 4.1: Number of experiments (01/2016 — 12/2018).

Table 4.1: Categories and its occurrences for likes, dislikes and recommendations.

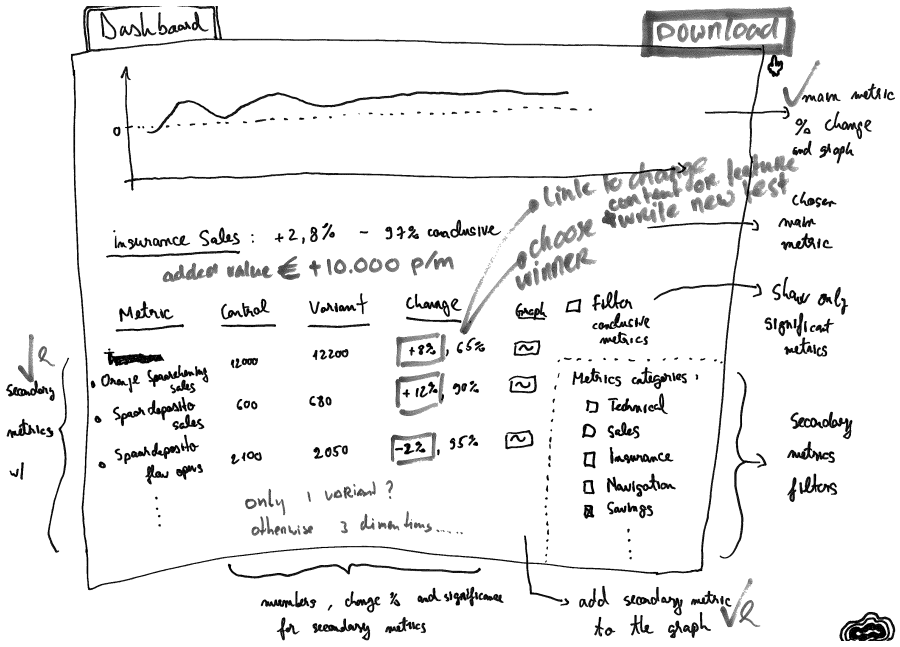
Positive		**Negative**		**Recommendation**	
Clear results	9	Limited results/metrics	12	Impact on experiments or business	6
Users are empowered to set up experiments	3	Limited testing	8	Select own metrics	5
Statistical significance	3	Creating test is hard	6	Notifications	5
Open platform	2	Missing database of earlier experiments	1	Visualisation of experiment	3
				Automated handling	3
				Database of earlier experiments	2
				User behaviour visualisation	2

4.2 Focus Group: Platform Good, Can Be Better

Thirteen people (all employees of the company) participated in the Focus Group session, of which 9 were Customer Journey Expert (CJE) and 4 Engineer (ENG). Overall the current experiment platform is good; *"Quick and easy to set up a test (if you know the instructions)."*, [p1, CJE] and *"Basics are there: randomiser, statistical significance, testing, APIs, persistence, etc."*, [$p3$, ENG]. However, users point out limited results/metrics as negative aspect; *"Not enough usefulness to be used extensively for business-oriented experiments."*, [$p4$, CJE], and *"Reporting is too simple. We need more information on significance, I want to see values."*, [$p7$, CJE]. All likes, dislikes and recommendations are categorised and their occurrences are counted (see Table 4.1).

During the discussion, most participants kept referring to the Ideal Dashboard (see Figure 4.2) as one of the participants had put down on paper during the brain-drawing part. In this brain-drawing, the most popular recommendations are reflected, such as impact on business, select own metrics and automated handling. According to the participant: *"Main metric which you choose yourself, for example, sales. Other metrics need to be included, with statistical significance testing. The advantage is that you really see what is going on in your application. Also, if your experiments are cannibalising other pages, this should be noted by the platform."*, [$p2$, ENG].

Figure 4.2: Brain-drawing of Ideal Dashboard in experiment platform.



This overview dashboard has selectable metrics, e.g. technical results or effect on sales. It will send notifications not only when the A/B test is completed, but also when the impact is negative such that users can react adequately.

Another interesting category for recommendation is database of earlier experiments. *"An ultimate overview dashboard of executed experiments, of the whole global company, where you have the option to filter experiments on team, page, conversion point or country. Bring your knowledge to all colleagues."*, [$p6$, CJE] . Unlike with the dashboard, not all participants agreed upon this feature. *"Yes, it will stimulate the test-cycle. It will prevent double testing; you upload a snap-shot such that you know and others can see what is been tested."*, [$p5$, CJE]. However, adding such an overview might not be all but positive: *"There is the risk of gamification, for example which teams are running most experiments at this moment? Teams will compete, and what are you going to do with the "worst performing team"?"*, [$p3$, ENG].

4.3 Survey: Experimentation Is Important, But Not Always Team's Focus

In total we received 295 responses, for a response rate of 20%, and a relatively low dropout rate; only 3% did not complete the survey. By discipline, 8.14% are Data Analysts (DA), 87.79% are Engineers (ENG), and 4.07% are Customer Journey Experts (CJE), with response rates of respectively 31.25%, 19.86% and 26.92%.

4.3.1 Categorisation

Coding the open questions resulted in a categorisation set. Answers are categorised using a decision tree (Figure 4.3). **Example:** "*When we do experiment, it's not about the little features because the effort doesn't weigh up to the value, we do however create usability and experience tests for new concepts to see if it fits*", [$p139$, ENG]. This person explains when experimentation is important, hence: Experimentation is important, Mentions Experiment Approach.

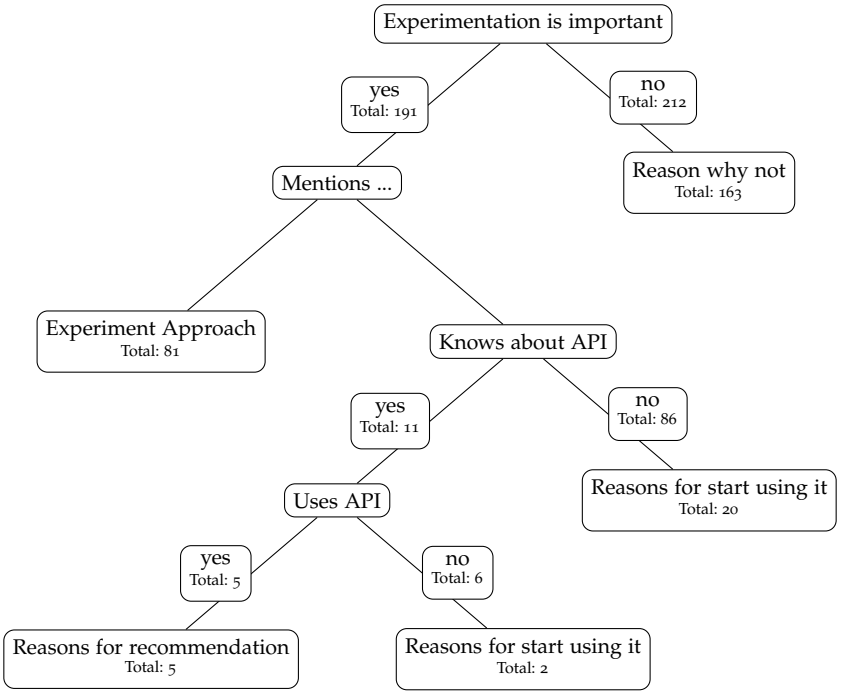


Figure 4.3: Decision tree for categorisation of survey responses on open questions.

Table 4.2: Questions with percentage of corresponding answer for the three different roles.

	Question	Response	Role CJE	DA	ENG
Q2	Have you heard of the experiment platform before?	Yes	16.7%	41.7%	24.3%
Q3	I wish my team would use experimentation more often to measure customer experience.	Agree	16.7%	25.0%	35.1%
Q4	At least two members of my team are passionate about experimentation.	Agree	0.0%	25.0%	24.7%
Q5	The experiment platform is easy to use.	Disagree	0.0%	8.3%	0.0%
Q7	What kind of education do you think is needed before you can use the experiment platform?	Documentation	25.0%	62.0%	60.2%

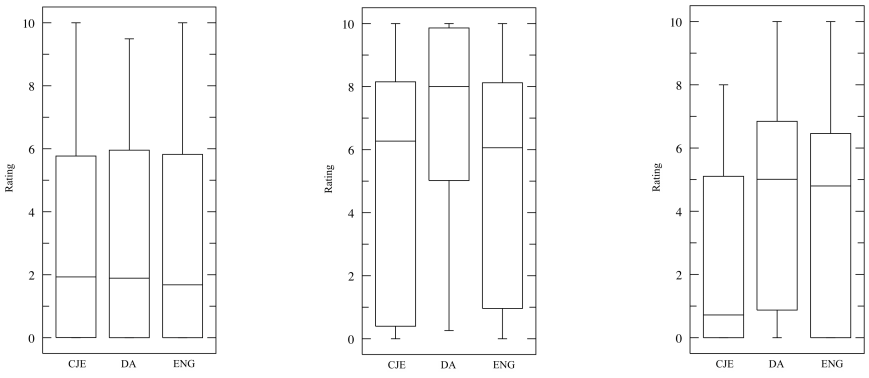
Rating differences by role (DA, ENG and CJE) using logistic regression models or the Shapiro-Wilk test identified no questions where having a specific role influenced the response: none of the questions have a significant answer and/or cross the correlation threshold.

4.3.2 Differences in Roles: Insignificant But Interesting

DA state that the experiment platform is not easy to use, whilst ENG and CJE do not use this response (see Table 4.2). Moreover, participants that are DA form a larger group that have heard of the experiment platform before, compared to ENG and CJE. According to DA and ENG responses, documentation is a sufficient approach for educating new users on the experiment platform. In second and third place are workshops and e-learnings.

The variation in the answers for the various roles is not only visible in the closed questions, but also in the slider bar questions (see Figure 4.4). All are unanimous when it comes to whether experimentation is part of a team's development cycle (Figure 4.4a). However, the answers are more diversified when it comes to recommending experimentation (Figure 4.4b) or the experiment platform in particular (Figure 4.4c).

DA are most likely to recommend experimentation. Reasons why are *"A/B testing is easy and proved to be successful in a number of occasions, but it is merely a "step-in" innovation. We, as ING, should aim higher!"*, [p27, DA] and *"I am advising colleagues to do so, especially if they can not find answers in the online data about customers"*, [p77, DA]. CJE are least likely to recommend the experiment platform: *"Getting fast feedback from clients is always useful, although experience shows that we don't have the proper client (test/experiment) base to select clients from and that we have dispersed teams, mine is actually doing IT work, almost never coming in contact with real clients"*, [p274, CJE], or simply because *"As I never heard of it and don't know what it is, it is very unlikely that I will recommend it"*, [p80, CJE].



(a) Boxplot for Q1[2] (b) Boxplot for Q8[3] (c) Boxplot for Q9[4]

Figure 4.4: Boxplots of slider bar questions.

[2] *Is experimentation part of your team's development cycle?*
[3] *How likely is it that you recommend using experimentation to a colleague?*
[4] *How likely is it that you recommend using the experiment platform to a colleague?*

4.4 Model: Pre-experiment Data Positively Influences Outcomes

For evaluating the influence of pre-experiment data the absolute effect size, power and p-value are considered. To learn whether the difference is statistically significant, the probability number retrieved from the test is compared against a pre-set threshold α. To know if an observed difference is not only statistically significant but also meaningful, it is important to calculate its effect size. In addition, a power analysis is conducted to ensure that the sample size is big enough for detecting differences.

4.4.1 Applied on Historical Experiments

Effect Size

Effect size is considered to be the difference between the two proportions [52]. The absolute effect size is calculated by $|\hat{\pi}_1 - \hat{\pi}_2|$. Each experiment has a small effect size without pre-experiment data. Most of the experiments have an effect size of .001 (41%), others .002 (25%), .01 (17%), or .0 (17%) (see Figure 4.5). Using pre-experiment data increases the effect size and thus the difference between the two proportions. Further investigation shows that, violating the assumption we made for the model, the conversion rate is unstable over time; it is on average 5.7 times larger in the pre-experiment period than during the experiment. Presumably, visitors reached the web page via an alternative route than via the A/B test. In this way, not all user behaviour is captured. Hence, the way traffic is tracked must be adjusted before pre-experiment data can be fully integrated. When measuring on events, instead of on a page (as in experiment 5), the characteristics of pre-experiment data comes closer to the data during the experiment.

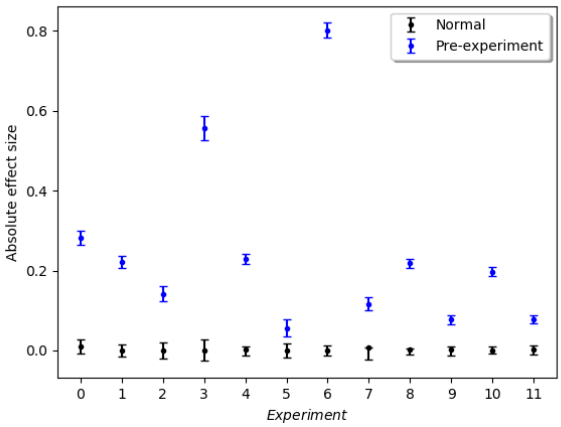


Figure 4.5: Effect size with confidence interval (=99%) for historical experiments with and without pre-experiment data.

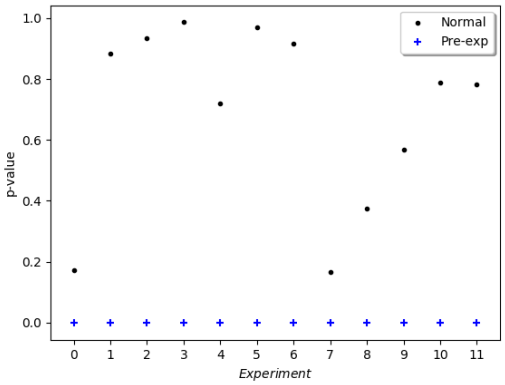
For all historical experiments, with the use of pre-experiment data, the p-value decreases close to 0.0, such that they are significant. However, the statistical power analysis (using estimated effect size) (see Figure 4.6b), shows that not all results are completely reliable. Naturally, a low power indicates a lower probability to detect a difference when it indeed exists and therefore indicates lower reliability. Even though experiment 5 has a low p-value when using pre-experiment data, namely $3.532e{-}11$, its estimated power does not cross the threshold of 0.8 [26]. All other experiments their power do cross this threshold, thus can be labeled with as adequate enough power.

Using pre-experiment data increases the effect size and thus also the difference between the two proportions for the historical experiments. Further investigation shows that, unlike our assumption when building the model, the conversion rate is unstable over time. For all historical experiments the conversion rate in the pre-experiment period is on average 5.7 times larger than during the experiment, which causes the effect size to increase. Moreover, for experiments 3 and 6, the number of conversions is higher than the number of visitors. This because each visitor at the web pages of ING is allowed to click a button multiple times. This causes the number of clicks to rise, whilst the number of visitors does not increase. Because of the lack of an extensive amount of experiments available at ING, we decided to simulate the model (see Section 4.4.2).

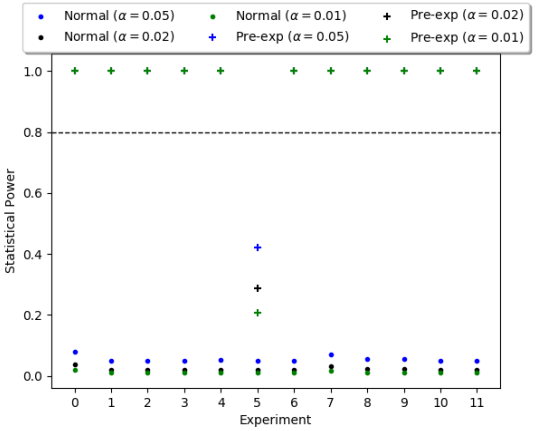
Power and p-value

Out of all the historical experiments in the experiment platform, only 12 had a begin and end date in the period of data collection, and had pre-experiment data available. For all these experiments the p-value decreases with the use of pre-experiment data, see Figure 4.6a. Moreover, all experiments have a p-value close to 0.0, such that the experiments are significant. However, when looked at results of the statistical power analysis (see Figure 4.6b), it becomes clear that not all results of the experiments are completely reliable. Naturally, a low power indicates a lower probability to detect a difference when it indeed exists, and therefore indicates a lower reliability. Even though experiment 5 has a low p-value when using pre-experiment data, namely 3.5326e−11, its estimated power does not cross the threshold of 0.8 [26]. All others do cross this threshold, thus can be labeled with enough power as statistically significant. Although the differences between 95%, 98% and 99% are minimal, it is best to choose 99% as confidence level. Currently, the experiment platform uses 95%, but it might be better to switch to 99% confidence level to make the calculation more accurate by reducing the number of Type I errors.

Figure 4.6: Power and p-value for historical experiments.



(a) P-value for historical experiments with and without pre-experiment data.



(b) Estimated power for historical experiments with and without pre-experiment data for three confidence levels.

Only experiment 5 does not cross the power threshold, this is due to the low amount of traffic. On average $n_1 = 8162$, $n_2 = 8176$ and $n_3 = 13\,459$ whereas for experiment 5 $n_1 = 1\,602$, $n_2 = 1\,592$, and $n_3 = 1\,198$.

4.4.2 Simulation

We assume an equal split of traffic, hence $n_1 = n_2$. Moreover, we assume that the pre-experiment version is equal to control, hence $\pi_3 = \pi_1$. π_1 varies from 0.1 to 0.9 with step size 0.1. The effect size is varied for each value of π_1. π_2 is calculated by π_1+ effect size. The effect sizes used are $0, 0.001, 0.002, 0.01, 0.02, 0.1$, and 0.2.

Number of Significant Tests

See Tables 4.3, 4.4, 4.5, 4.6, 4.7, 4.8, 4.9 for the number of significant tests out of 1 000 simulations for different sizes of n_1, n_2 and n_3, and $\alpha = 0.05$ and $\alpha = 0.01$. π_1 is set to 0.1, since at ING on average $\pi_1 = 0.1211$. Now, unlike at ING, the assumption $\pi_1 = \pi_3$ holds. From these results we can conclude that the model for a situation without any effect size (Table 4.3) works as expected; for $\alpha = 0.05$ on average 5.5% and for $\alpha = 0.01$ on average 1.6% of the 1 000 simulations are significant (i.e., these are Type I errors). Overall, using pre-experiment data is for 26 of the 36 cases favourable for the amount of significant tests; it increases and is higher than without pre-experiment data. Rerunning the experiments for the results in Table 4.3 shows similar trends, for example for $n_3 = 10\,000$ the highest value is retrieved for $n_3 = 10$ and the lowest values are retrieved when n_3 equals 100 or 1 000. Nevertheless, randomisation and potentially the amount of simulations cause small variance in the amount of significant tests; a maximum variation of 1.5% and an average variation of 0.6% in significant tests.

Table 4.3: Number of significant tests out of 1 000 simulations for effect size = 0, ($\pi_1 = 0.1$, $\pi_2 = 0.1$).

$n_3/n_1 = n_2$	100		1000		10 000	
	0.05	0.01	0.05	0.01	0.05	0.01
0	50	15	59	15	55	17
10	52	17	59	15	55	18
100	41	5	44	8	52	14
1000	49	8	49	7	51	14

The percentage of significant tests increases as the effect size increases. For smaller effect sizes (0.001 and 0.002), the differences between the different sizes of pre-experiment data are minimal. However, the larger the amount of visitors for version A and B ($n_1 = n_2$), the larger the number of significant tests. This is especially the case for $n_3 = 100$ and effect size 0.001 (Table 4.4 and 4.5). This is also the case for effect sizes 0.01 (Table 4.6) and 0.02 (Table 4.7).

Most of the A/B tests at ING (used in Section 4.4), have an effect size of 0.001 (41%), others have an effect size of 0.002 (25%), or 0.000 (17%), or $\sim$ 0.01 (17%). The average value for $n_1 = n_2 = 8651$, hence we use $n_1 = n_2 = 10\,000$ as indicator for the experiments at ING. According to the simulation, choosing $n_3 = 100$ as the amount of pre-experiment data provides the highest percentage of significant tests for an effect size of 0.001, independent of the chosen value for significant level α. For an effect size of 0.002, the highest percentage is retrieved by using no pre-experiment data for $\alpha = 0.05$. However, for $\alpha = 0.01$ the highest percentage is retrieved by using the maximum amount of pre-experiment data ($n_3 = 1\,000$). For an effect size of 0.01, the maximum percentage of significant tests is retrieved by using the maximum amount of pre-experiment data ($n_3 = 1\,000$) and maximum amount of traffic for both versions ($n_1 = n_2 = 10\,000$). This is not the same for effect size of 0.02, here works $n_3 = 100$ the best.

Table 4.4: Number of significant tests out of 1 000 simulations for effect size = 0.001, ($\pi_1 = 0.1$, $\pi_2 = 0.101$).

$n_1 = n_2$	100		1000		10 000	
n_3	0.05	0.01	0.05	0.01	0.05	0.01
0	53	7	48	10	69	15
10	45	6	47	11	69	16
100	53	10	46	12	95	20
1000	58	9	48	6	61	11

Table 4.5: Number of significant tests out of 1 000 simulations for effect size = 0.002, ($\pi_1 = 0.1$, $\pi_2 = 0.102$).

$n_1 = n_2$	100		1000		10 000	
n_3	0.05	0.01	0.05	0.01	0.05	0.01
0	48	11	46	5	86	19
10	42	9	43	6	85	19
100	40	8	58	14	71	17
1000	51	9	53	13	83	21

Table 4.6: Number of significant tests out of 1 000 simulations for effect size = 0.01, ($\pi_1 = 0.1$, $\pi_2 = 0.11$).

$n_1 = n_2$	100		1000		10 000	
n_3	0.05	0.01	0.05	0.01	0.05	0.01
0	57	6	108	32	637	384
10	54	11	106	32	638	392
100	59	19	126	32	618	376
1000	70	19	127	41	652	438

Table 4.7: Number of significant tests out of 1 000 simulations for effect size = 0.02, ($\pi_1 = 0.1$, $\pi_2 = 0.12$).

$n_1 = n_2$	100		1000		10 000	
n_3	0.05	0.01	0.05	0.01	0.05	0.01
0	68	13	294	114	993	977
10	69	14	302	115	993	977
100	87	25	303	122	998	983
1000	118	40	407	201	993	977

Table 4.8: Number of significant tests out of 1 000 simulations for effect size = 0.1, ($\pi_1 = 0.1$, $\pi_2 = 0.2$).

$n_1 = n_2$	100		1000		10 000	
n_3	0.05	0.01	0.05	0.01	0.05	0.01
0	499	296	1 000	1 000	1 000	1 000
10	534	294	1 000	1 000	1 000	1 000
100	656	426	1 000	1 000	1 000	1 000
1000	834	659	1 000	1 000	1 000	1 000

Table 4.9: Number of significant tests out of 1 000 simulations for effect size = 0.2, ($\pi_1 = 0.1$, $\pi_2 = 0.3$).

$n_1 = n_2$	100		1000		10 000	
n_3	0.05	0.01	0.05	0.01	0.05	0.01
0	696	873	1 000	1 000	1 000	1 000
10	973	898	1 000	1 000	1 000	1 000
100	994	961	1 000	1 000	1 000	1 000
1000	999	994	1 000	1 000	1 000	1 000

Power

See Table 4.10 for the power calculations for different values for n_1, n_2 and n_3, various effect sizes and two significance levels. From these results it becomes clear that the power for effect sizes smaller than 0.1 do not cross the 0.8 threshold, which means that the power is too low for these cases. The sample sizes used for these cases is too small. See Table 4.11 for the sample sizes needed, in case of $n_3 = 0$ and the assumption $n_1 = n_2$.

Although the differences in power level are minimal, Figures 4.7 and 4.7 show that the amount of pre-experiment data does influence the power level. The power analysis of the simulation shows that the model is most reliable for $n_1 = n_2 = 10\,000$, and works with enough statistical power for effect sizes 0.01 and 0.02.

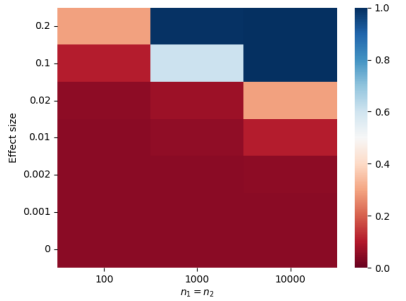
Table 4.10: Power calculation for simulation.

	$n_1 = n_2$	100		1 000		10 000	
n_3	effect size	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
0	0	0.05	0.01	0.05	0.01	0.05	0.01
	0.001	0.05000573	0.01000341	0.05005728	0.01001863	0.05057295	0.01018651
	0.002	0.05002291	0.01000745	0.05022913	0.01007454	0.0522942	0.01074942
	0.01	0.05057295	0.01018651	0.05574725	0.0118904	0.1089546	0.03134437
	0.02	0.0522942	0.01074942	0.07320971	0.01789427	0.2929889	0.1227288
	0.1	0.1089546	0.03134437	0.6087795	0.3670189	0.9999998	0.9999965
	0.2	0.2929889	0.1227288	0.9940005	0.9710402	1	1
10	0	0.05	0.01	0.05	0.01	0.05	0.01
	0.001	0.050006	0.01000341	0.05005756	0.01001872	0.05057324	0.0101866
	0.002	0.050024	0.0100078	0.05023027	0.01007491	0.05229534	0.0107498
	0.01	0.05060025	0.01019541	0.05577594	0.01189995	0.1089848	0.03135635
	0.02	0.0524036	0.01078533	0.07332658	0.01793574	0.2931099	0.1228006
	0.1	0.1118294	0.03249115	0.6109109	0.3691128	0.9999999	0.9999966
	0.2	0.3044888	0.1296178	0.9941868	0.9717667	1	1
100	0	0.05	0.01	0.05	0.01	0.05	0.01
	0.001	0.05000764	0.01000341	0.05006	0.01001951	0.0505758	0.01018744
	0.002	0.05003055	0.01000993	0.05024005	0.01007809	0.05230563	0.01075317
	0.01	0.05076402	0.01024881	0.0560219	0.01198182	0.1092547	0.03146365
	0.02	0.05306033	0.01100123	0.07432878	0.01829206	0.2941932	0.1234437
	0.1	0.12917	0.03960702	0.6288288	0.3870028	0.9999999	0.9999968
	0.2	0.3720084	0.1728952	0.99557	0.9773333	1	1
1000	0	0.05	0.01	0.05	0.01	0.05	0.01
	0.001	0.0500105	0.01000341	0.05007637	0.01002484	0.05060025	0.01019541
	0.002	0.050042	0.01001366	0.05030552	0.0100994	0.0524036	0.01078533
	0.01	0.05105072	0.01034237	0.05767157	0.01253297	0.1118294	0.03249115
	0.02	0.05421084	0.01138081	0.08106697	0.02071976	0.3044888	0.1296178
	0.1	0.159807	0.05299321	0.73304	0.5024574	0.9999999	0.9999984
	0.2	0.4820633	0.2543177	0.9993224	0.9951753	1	1

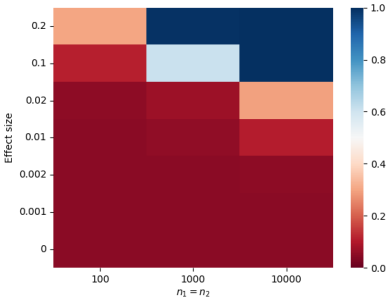
Table 4.11: Sample size calculations for pre-set effect size and power level.

	power = 80%		power = 85%		power = 90%		power = 95%	
effect size	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
0.001	15 697 721	23 357 936	1 7956 776	26 096 884	21 014 839	29 758 774	25 989 418	35 628 329
0.002	3 924 430	5 839 484	4 489 194	6 524 221	5 253 710	7 439 694	6 497 355	8 907 082
0.01	156 977	233 579	179 568	260 969	210 148	297 588	259 894	356 283
0.02	39 244	58 395	44 892	65 242	52 537	74 397	64 974	89 071
0.1	1 570	2 336	1 796	2 610	2 101	2 976	2 599	3 563
0.2	392	584	449	652	525	744	650	891

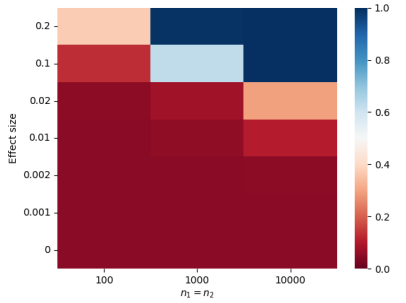
Figure 4.7: Power for simulation with confidence level 95%.



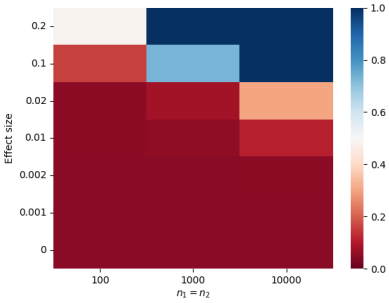
(a) $n_3 = 0$



(b) $n_3 = 10$

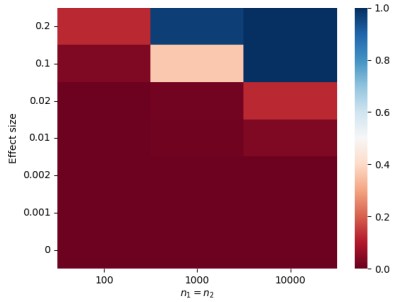


(c) $n_3 = 100$

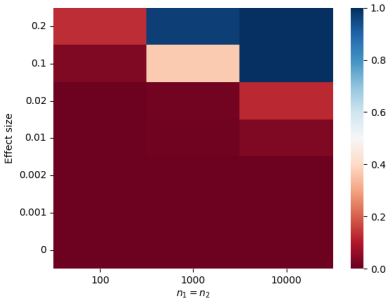


(d) $n_3 = 1\,000$

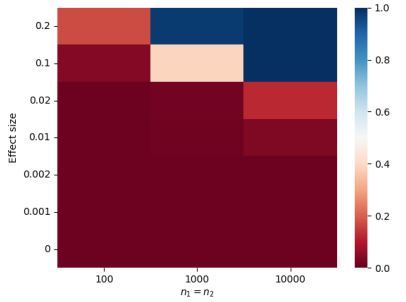
Figure 4.8: Power for simulation with confidence level 99%.



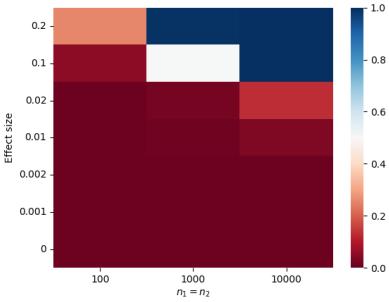
(a) $n_3 = 0$



(b) $n_3 = 10$



(c) $n_3 = 100$



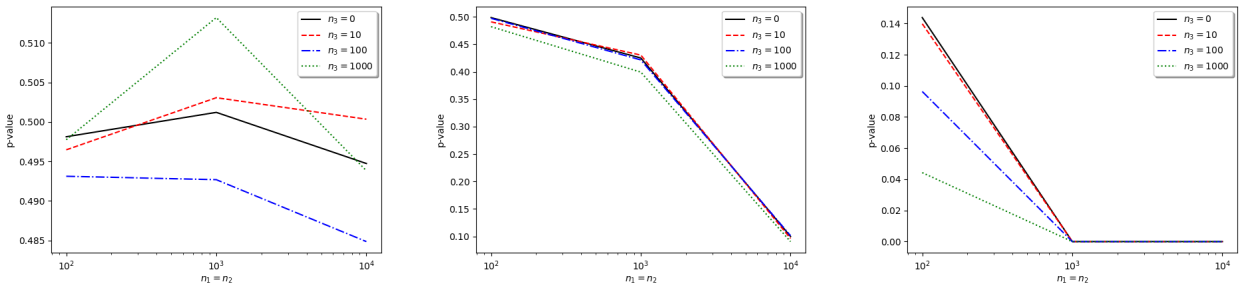
(d) $n_3 = 1\,000$

P-Value

Increasing the amount of pre-experiment data does not massively influence the mean and standard deviation of the p-value. However, as the effect size increases the p-value does become smaller. Also, the standard deviation of the p-value decreases, which means that the points are closer to the mean. Hence, a larger effect size means that conclusions can be drawn with more certainty (a lower value for p). This is in line with the power analysis in Section 4.4.2, where the power increases as the effect size increases. Thus, more powerful tests tend to generate smaller p-values.

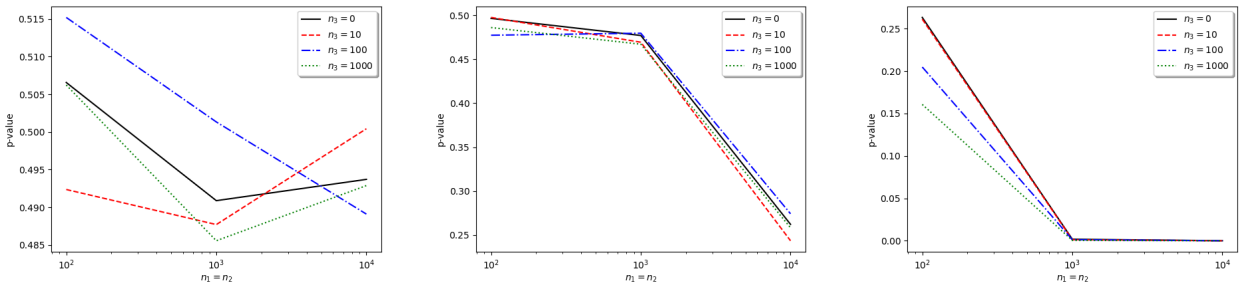
The trends for the three effect sizes (0.001, 0.01 and 0.1) are similar for $\pi_1 = 0.1$ and $\pi_1 = 0.5$ (see Figure 4.9 and 4.10). The most stable situation arises when the amount of pre-experiment data is small; for each situation it results in one of the lowest p-values, regardless of the amount of visitors ($n_1 = n_2$). However, choosing a larger amount of pre-experiment data for effect size 0.1 results in a smaller p-value, namely on average 0.143813 for $n_3 = 0$, 0.139870 for $n_3 = 10$, 0.096365 for $n_3 = 100$, and 0.044273 for $n_3 = 1000$. The effect of pre-experiment data is for a small sample size is arguable. Figures 4.9a and 4.10a show a bit of chaos, compared to the other figures. This variance and uncertainty is due to the low power as represented in Table 4.10 (Section 4.4.2). From the sample size calculations it becomes clear that the sample sizes used are at least 16K times too small, to retrieve at least a power level of 80%.

Figure 4.9: Average p-value for different effect sizes for $\pi_1 = 0.1$.



(a) effect size = 0.001 (b) effect size = 0.01 (c) effect size = 0.1

Figure 4.10: Average p-value for different effect sizes for $\pi_1 = 0.5$.



(a) effect size = 0.001 (b) effect size = 0.01 (c) effect size = 0.1

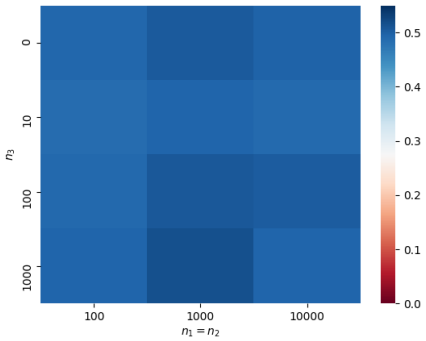
Patterns in P-Value and Effect Size

In order to find patterns in p-values in relation to effect size, for all sizes of traffic and pre-experiment data the p-values are averaged for three different click-through rates $(0.1, 0.5, 0.8)$. See Figures 4.11, 4.12, 4.13. These heat maps show that as the effect size increases the p-value decreases. Also, the p-values for the click-through rates show a similar trend for each effect size.

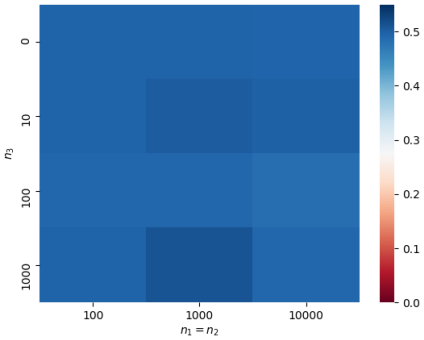
Interestingly, a click-through rate of 0.1 and 0.8 show a more similar pattern, than 0.5 does with the these two click-through rates. The largest differences in p-values across the different click-through rates can be found in the smaller effect sizes, respectively 0.01 and 0.02. Although the the values for the p-values are all in the same area, namely between 0.03205 and 0.472, the heat maps do not show as much resemblance as the heat maps for the other effect sizes. Also, these effect sizes show the largest distribution of p-values. For instance, for effect size 0.02, and $n_1 = n_2 = 1000$ the p-value is small enough to make the test significant, whilst for $n_1 = n_2 = 100$ this is not the case.

All in all, using more pre-experiment data is not always necessarily better. For small effect sizes (0.001 or 0.002) in most cases in terms of pre-experiment data "less is more" (either $n_3 = 10$ or $n_3 = 100$), resulting in a lower p-value than if no or the maximum amount of pre-experiment data was used. This is in line with what Deng et. al. found. According to them the higher the correlation, the better the variance reduction. "Increasing the pre-experiment period increases coverage because of having a better chance of matching an experiment user in the pre-experiment period." However, there is a constraint to the amount of pre-experiment data: "Using a pre-experiment period of 1-2 weeks works well for variance reduction. Too short a period will lead to poor matching, whereas too long a period will reduce correlation with the outcome metric during the experiment period." For a larger effect size of 0.1 or 0.2 using pre-experiment data does not influence the amount of significant tests, as all p-values are below the thresholds for $\alpha = 0.05$ and $\alpha = 0.01$. Nevertheless, in most cases using the maximum amount of pre-experiment data ($n_3 = 1\,000$) leads to the lowest p-values for the largest traffic numbers.

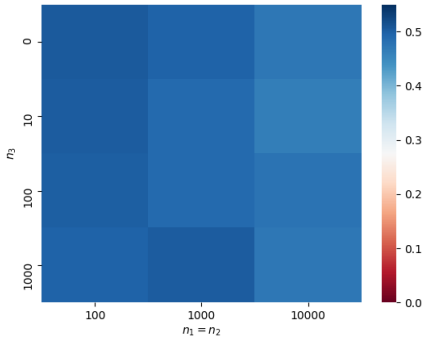
Figure 4.11: Average p-value for $\pi_1 = 0.1$ and different effect sizes.



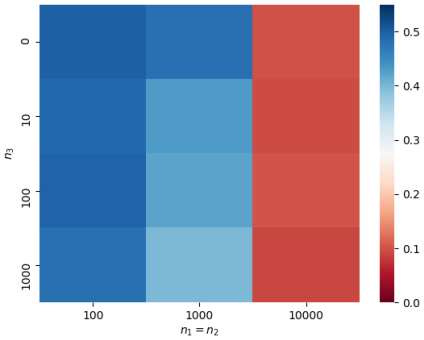
(a) Effect size = 0



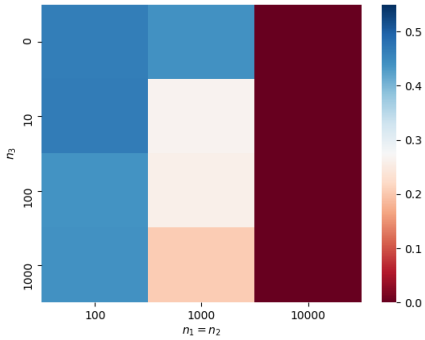
(b) Effect size = 0.001



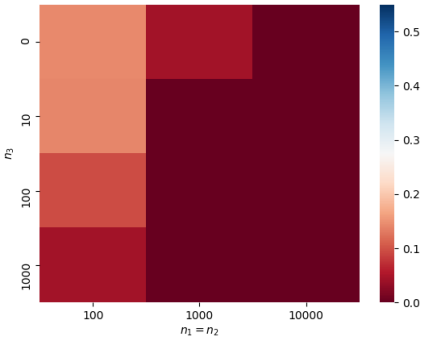
(c) Effect size = 0.002



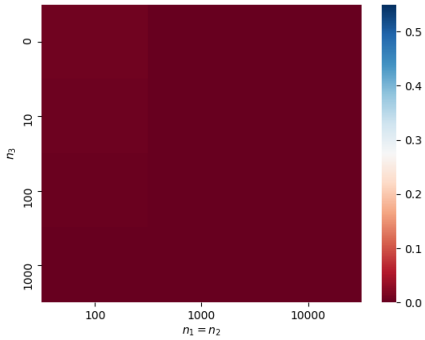
(d) Effect size = 0.01



(e) Effect size = 0.02

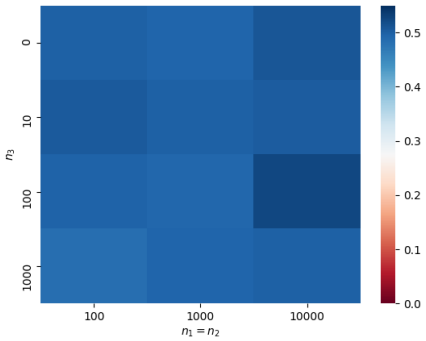


(f) Effect size = 0.1

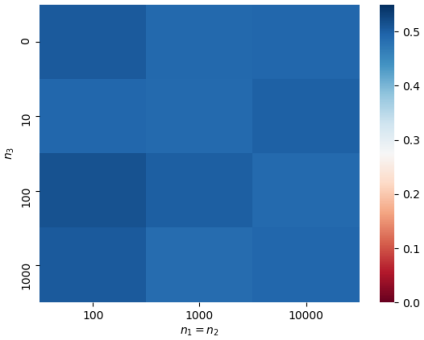


(g) Effect size = 0.2

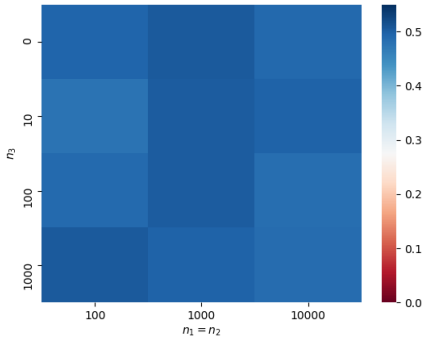
Figure 4.12: Average p-value for $\pi_1 = 0.5$ and different effect sizes.



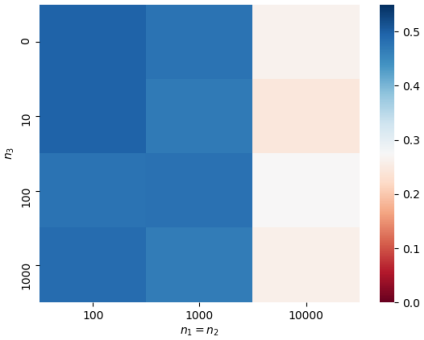
(a) Effect size = 0



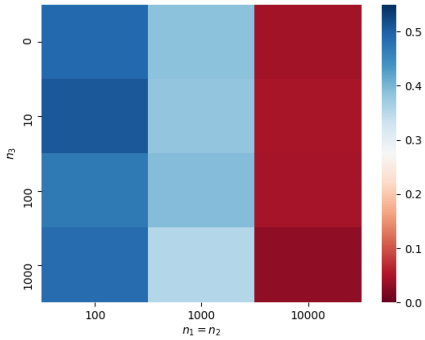
(b) Effect size = 0.001



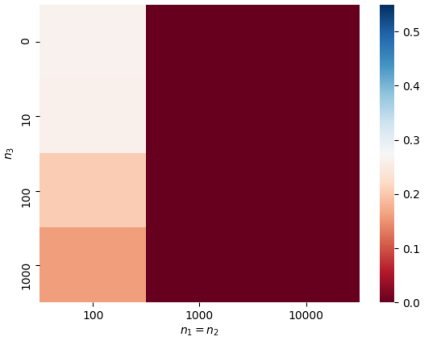
(c) Effect size = 0.002



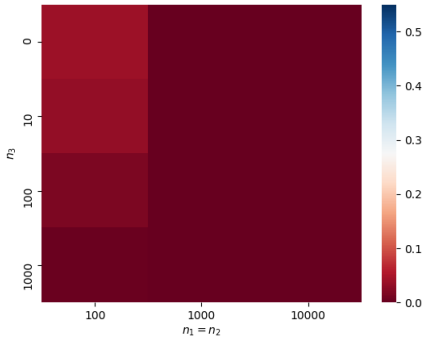
(d) Effect size = 0.01



(e) Effect size = 0.02

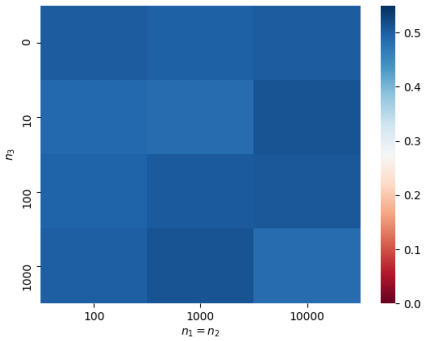


(f) Effect size = 0.1

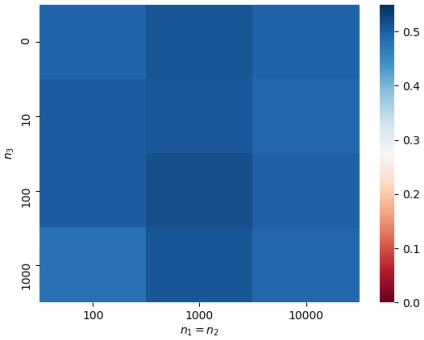


(g) Effect size = 0.2

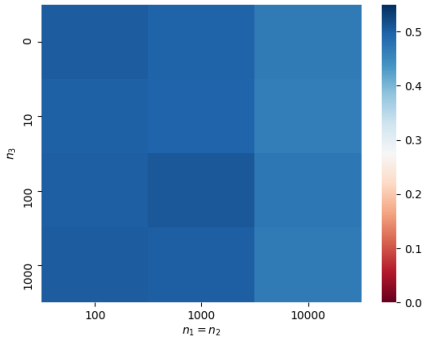
Figure 4.13: Average p-value for $\pi_1 = 0.8$ and different effect sizes.



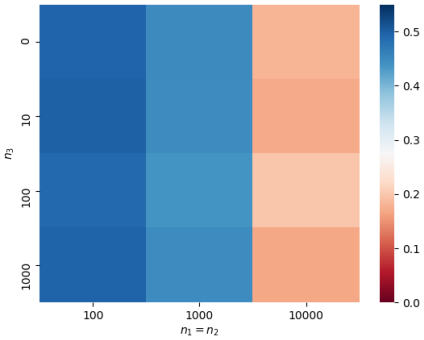
(a) Effect size = 0



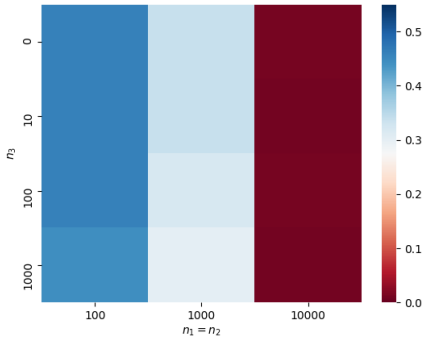
(b) Effect size = 0.001



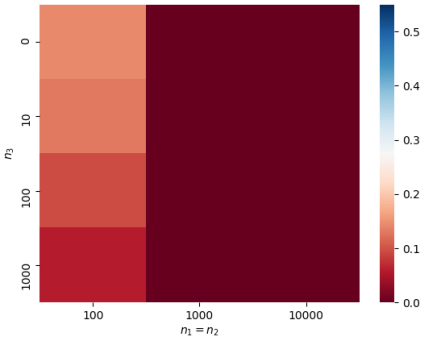
(c) Effect size = 0.002



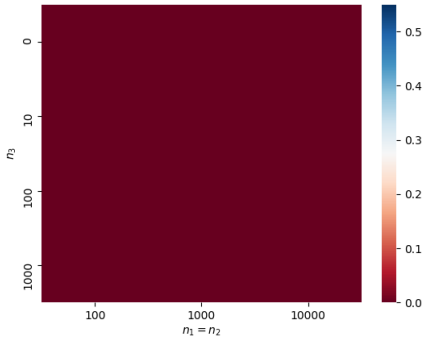
(d) Effect size = 0.01



(e) Effect size = 0.02



(f) Effect size = 0.1



(g) Effect size = 0.2

Chapter 5

Discussion and Conclusions

5.1 Discussion

In this section, the research questions as described in Section 1.2 are answered.

> *RQ 1: What factors influence the usage of continuous experimentation in cross-functional teams?*

Current users state that the experiment platform helps to verify the usability of new concepts. According to them, the experiment platform is easy to use (given that one knows how to set up a test), and all basic metrics are there. The majority of the participants had not heard of the experiment platform before filling in the survey, which could be one of the reasons why teams are not using it. Another reason that also came forward out of the survey is that experimentation, in general, is not part of a team's focus. In some cases this is because of the lack of time for experimentation, in other cases teams work on applications for internal use only. Nevertheless, participants do wish that experimentation would be used more often to measure customer experience. The majority agrees that documentation is a good form of education for new users. Current users do find the results clear and like the fact that they can prove ideas with facts using an open platform. The experiment platform can be improved by providing notifications and automated handling of completed experiments. By providing selectable metrics, including the impact on business, the results of an experiment will be less limited and users are more likely to use the experiment platform more often.

> *RQ 2: What can be done to increase the power and sensitivity of A/B test outcomes?*

Pre-experiment data positively influences the results of algorithms; more A/B tests are significant, and their power increases. For the historical experiments at ING, using pre-experiment data increases the effect size and thus also the difference between the two proportions. Further investigation shows that, unlike our assumption when building the model, the conversion rate is unstable over time. Moreover, the data used as pre-experiment data is noisy, causing the results to be too optimistic. In summary, it is not possible to integrate our proposed model into ING's experimentation platform without making changes to the platform first. According to our simulation the model works best for a large

amount of traffic and an effect size of at least 0.1. The power for very small effect sizes (0.001) is very low. In order to increase the number of significant tests, the experiment platform could make an estimation of the effect of an A/B test before running it to determine the minimum sample size needed based on information (such as effect size, significance level, and desired power). In that way, users will know whether a test makes sense prior to execution such that time and effort can be saved.

Not only does using pre-experiment data increase the number of significant A/B test outcomes, but it also ensures better interpretation and prediction of user behaviour. Applications for this could be for example on a website which sells products to customers. **Example:** If the Click-Through-Rate for a web page which invites potential customers for a mortgage meeting (such as in Figure 3.2) could be increased by 1%, this means extra potential customers, and thus extra (potential) revenue. In numbers this means, even if the effect size is rather small, for example, equal to 0.001, and this web page receives one million visitors ($n_1 = 1\,000\,000$). Then, the appointments made increases with $0.001 * 1\,000\,000 = 1\,000$ extra potential customers.

This study has some limitations within which the presented findings need to be carefully interpreted. First, outcomes of empirical studies are never proof. It can only support a hypothesis, reject it, or do neither. Therefore, this study needs to be interpreted as such. Secondly, bias and inaccurate responses influence the interpretation of the survey results. To limit the consequences, established guidelines for designing, executing and analysing a survey were followed. The danger of a focus group is group consensus; to prevent this from happening participants were asked to first write down their ideas before discussing with others, such that also the individual opinions were captured. Thirdly, the experiment platform has been in use for two years, this might not be enough to say something about adoption of this technology. The majority could be early adopters, but this can only be stated with more insights in the technology life cycle [45] of the platform. Lastly, single case study analysis has a few limitations as well, such as researcher subjectivity and external validity. In order to minimise these limitations, work has been executed collaboratively with Universities and future replication of this work in other organisations will indicate whether or not results can be generalised.

5.1.1 Future Work

Another way to improve outcomes of A/B tests is to change the way traffic is handled. Google Analytics uses a multi-armed bandit (MAB) approach to managing online experiments [46]. The main idea of this approach is to explore and exploit settings at the same time. Part of the traffic is equally split over the variants, whilst the rest of the traffic is sent to the best performing variant [19]. Although the average conversion rate will be higher using this methodology, it is more difficult to calculate statistical differences in case of little traffic. The main benefit for using MAB is that traffic is handled more effectively, such that a larger part of your traffic can be sent through experimentation. MAB is mostly suited for continuous experimentation where the focus lies on increasing conversion rates [10].

"Peeking" [32] at data regularly and stopping an experiment as soon as the preliminary results are convincing enough to draw a conclusion leads to an increase of Type I errors, or false positives. Although the experiment might seem successful, the resulting significance tests are invalid [9]. Walsh et al. [12] state that traditional p-values and confidence intervals give unreliable interference when users peek. Instead, presenting robust data allows users to draw more reliable conclusions regarding running experiments.

As stated in Section 3.5.1, multivariate testing is not available in the experiment platform. Various current users stated that testing multiple variants at the same time is desired, but unfortunately not supported by the current experiment platform. Both the advantages and disadvantages should be taken into consideration before implementing this feature. The main advantage [18] of multivariate testing is that it limits the amount of sequentially run A/B tests. However, having more changes to test, the biggest limitation is amount of traffic needed to complete the test. Also, analysing test results for a multivariate test is more complex than analysing test results for an A/B test.

5.2 Conclusions

In conclusion, stimulating the adoption of A/B testing using the experiment platform depends on several factors. Firstly, teams should be informed about the existence and possibilities of the experiment platform. Education in the form of strong documentation is sufficient. Secondly, in order to empower current users to increase their usage, the experiment platform could be improved on two levels; usability and results. By providing notifications and automated handling, processing results will be easier. When users can choose their own metrics, users have more freedom to adapt the experiments to their needs. Secondly, A/B tests become more sensitive by increasing the confidence level from 95% to 99% and, their power increases by expanding significance calculations with pre-experiment data. Pre-experiment data decreases the p-value and its standard deviation. A larger effect size helps to draw conclusions with more certainty. More powerful tests tend to generate smaller p-values. The use of pre-experiment data comes with a few limitations; Pre-experiment data is only useful when the conversion ratio is stable over time. Also, the pre-experiment model works best for an effect size larger than currently is used for A/B tests at ING, hence the pre-experiment model will only work if larger differences will be tested at ING, or tests run for a longer time period to increase the amount of traffic to a desired level. Thirdly, preventing users to perform A/B tests with too little power or too little traffic by providing a test during the set-up of their experiments, will help to save time and effort. Implementing all previously described improvements potentially increases engagement of A/B testing at ING. Knowledge sharing in the research field of A/B testing is dominated by practical guides in the form of blogs. The academic world should invest more in empirical research to keep up with the developments at companies in the practical field.

Acknowledgements

I wish to thank various people for their contribution to this project. I would like to express my very great appreciation to Matthijs van Leeuwen, my research supervisor, for his patient guidance, enthusiastic encouragement and useful critiques and recommendations throughout the project. I would like to extend my thanks to my co-supervisor Xishu Li for her adequate support and constructive suggestions. The internship opportunity I had with ING Netherlands was a great chance for learning and my professional development. I am also grateful for having a chance to work with and learn from so many people and professionals, specifically Kevin Anderson, Arie van Deursen, and Hennie Huijgens. Finally, I would like to thank my parents for their unconditional support, love, and encouragement throughout my life.

Bibliography

[1] Ling 300. Tutorial: Pearson's chi-square test for independence. "https://www.ling.upenn.edu/~clight/chisquared.htm", 2008. [Accessed 06-12-2018].

[2] Pervaiz K Ahmed and Mohammed Rafiq. Internal marketing issues and challenges. *European Journal of marketing*, 37(9):1177–1186, 2003.

[3] Alan Agresti, Barbara Finlay. *Statistical Methods for the Social Sciences (4th Edition)*. Pearson Education Limited, Essex, 2014.

[4] Alex Deng, Ya Xu , Ron Kohavi, Toby Walker. Wsdm '13 proceedings of the sixth acm international conference on web search and data mining. In *Improving the Sensitivity of Online Controlled Experiments by Utilizing Pre-Experiment Data*, pages 123–132, New York, 2013. ACM.

[5] Alex Deng, Ya Xu , Ron Kohavi, Toby Walker. Proceeding kdd '16 proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. In *Improving the Sensitivity of Online Controlled Experiments: Case Studies at Netflix*, pages 645–654, New York, 2016. ACM.

[6] Andrew Begel and Thomas Zimmermann. Analyze this! 145 questions for data scientists in software engineering. In *Proceedings of the 36th International Conference on Software Engineering*, pages 12–23. ACM, 2014.

[7] Netflix Technology Blog. A/b testing and beyond: Improving the netflix streaming experience with experimentation and data science. "https://medium.com/netflix-techblog/a-b-testing-and-beyond-improving-the-netflix-streaming-experience-with-experimentation-and-data-5b0ae9295bdf", 2017. [Accessed 04-12-2018].

[8] Aleksander Fabijan , PAvel Dmitriev, Helena Holmström Olsson , Jan Bosch. The evolution of continuous experimentation in software product development. *International Conference on Software Engineering*, 39th, 2017.

[9] Callie McRee and Kelly Shen. How etsy handles peeking in a/b testing. "https://codeascraft.com/2018/10/03/how-etsy-handles-peeking-in-a-b-testing/", 2018. [Accessed 14-11-2018].

[10] Paras Chopra. Why multi-armed bandit algorithm is not better than a/b testing. "https://vwo.com/blog/multi-armed-bandit-algorithm/", 2012. [Accessed 30-10-2018].

[11] The SciPy community. *scipy.stats.norm*. SciPy, 2019. SciPy version 1.2.1.

[12] David Walsh, Ramesh Johari, Leonid Pekelis. Kdd '17 proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining. In *Peeking at A/B Tests: Why it matters, and what to do about it*, pages 1517–1525, New York, 2017. ACM.

[13] Deloitte University Press. Rewriting the rules for the digital age. "https://www2.deloitte.com/content/dam/Deloitte/lu/Documents/human-capital/lu-hc-2017-global-human-capital-trends-gx.pdf|", December 2017. [Accessed 24-10-2018].

[14] Diane Tang, Ashish Agarwal, Deirdre O'Brien, Mike Meyer. Proceedings of the 16th acm sigkdd international conference on knowledge discovery and data mining, 2010.

[15] Eleri Dixon, Emily Enos, Scott Brodmerkle. A/B testing of a webpage. https://patents.google.com/patent/US7975000B2/en, 01 2005. [Accessed 04-10-2018].

[16] Aleksander Fabijan, Pavel Dmitriev, Helena Holmstrom Olsson, and Jan Bosch. Online controlled experimentation at scale: An empirical survey on the current state of a/b testing. In *2018 44th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, pages 68–72. IEEE, 2018.

[17] Murray J Fisher, Andrea P Marshall, and Marion Mitchell. Testing differences in proportions. *Australian Critical Care*, 24(2):133–138, 2011.

[18] Hanneke Verhoef. Dis- and advantages of multivariate tests. "https://www.abtasty.com/nl/blog/voor-en-nadelen-van-multivariate-tests/". [Accessed 14-03-2019].

[19] Steve Hanov. 20 lines of code that will beat a/b testing every time. "http://stevehanov.ca/blog/index.php?id=132", 2012. [Accessed 30-10-2018].

[20] Ronggui HUANG. *RQDA: R-based Qualitative Data Analysis*. RQDA, 2017. R package version 0.3-0.

[21] ING. Ing group annual report 2017, 2017.

[22] Simon Jackson. How booking.com increases the power of online experiments with cuped. "https://booking.ai/how-booking-com-increases-the-power-of-online-experiments-with-cuped-995d186fff1d", 2018. [Accessed 04-12-2018].

[23] Amit Kedia. Advantages and disadvantages of custom software / application. "https://www.linkedin.com/pulse/advantages-disadvantages-custom-software-application-amit-kedia/", 2015. [Accessed 25-10-2018].

[24] Barbara Kitchenham, Lech Madeyski, David Budgen, Jacky Keung, Pearl Brereton, Stuart Charters, Shirley Gibbs, and Amnart Pohthong. Robust statistical methods for empirical software engineering. *Empirical Software Engineering*, 22(2):579–630, 2017.

[25] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M Henne. Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery*, 18(1):140–181, 2009.

[26] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M Henne. Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery*, 18(1):140–181, 2009.

[27] Lars Backstrom, Jure Leskovec. Network bucket testing. "https://research.fb.com/publications/network-bucket-testing/", 2011. [Accessed 04-12-2018].

[28] Valentina Lenarduzzi and Davide Taibi. Mvp explained: A systematic mapping study on the definitions of minimal viable product. In *Software Engineering and Advanced Applications (SEAA), 2016 42th Euromicro Conference on*, pages 112–119. IEEE, 2016.

[29] Ron Kohavi , Roger Longbotham. *Online Controlled Experiments and A/B Testing*. Springer Science+Business Media New York, Encyclopedia of Machine Learning and Data Mining, 2017.

[30] Sezin Gizem Yaman, Myriam Munezero, Jürgen Münch, Fabian Fagerholm , Ossi Syd , Mika Aaltola , Christina Palmu , Tomi Männistöa. Introducing continuous experimentation in large software-intensive product and service organisations. *The Journal of Systems and Software*, 133:195–211, 2017.

[31] Martyn Shuttleworth. Case study research design. "https://explorable.com/case-study-research-design", 2008. [Accessed 26-11-2018].

[32] Evan Miller. How not to run an a/b test. "http://www.evanmiller.org/how-not-to-run-an-ab-test.html", 2010. [Accessed 14-11-2018].

[33] Janice M Morse, Michael Barrett, Maria Mayan, Karin Olson, and Jude Spiers. Verification strategies for establishing reliability and validity in qualitative research. *International journal of qualitative methods*, 1(2):13–22, 2002.

[34] Fabian Fagerholm, Alejandro Sanchez Guinea , Hanna Mäenpää, Jürgen Münch. The right model for continuous experimentation. *The Journal of Systems and Software*, 123:292–305, 2017.

[35] Science Museum. James lind (1716-94)p. "http://broughttolife.sciencemuseum.org.uk/broughttolife/people/jameslind", October 2018. [Accessed 04-10-2018].

[36] Patricia Lotich. What is the purpose and advantages of focus group interviews? "https://www.socialmediatoday.com/content/what-purpose-and-advantages-focus-group-interviews", 2011. [Accessed 31-10-2018].

[37] Robin L Plackett. Karl pearson and the chi-squared test. *International Statistical Review/Revue Internationale de Statistique*, pages 59–72, 1983.

[38] Fatemeh Rabiee. Focus-group interview and data analysis. *Proceedings of the nutrition society*, 63(4):655–660, 2004.

[39] Raphael Lopez Kaufman ,Jegar Pitchforth, Lukas Vermeer. Presented at the 2017 conference on digital experimentation (code@mit), 2017.

[40] Richard F. Ramsey. Generating Cohen's Effect Size ḧVia Arcsin Arcsine Transformations. "https://people.ucalgary.ca/~ramsay/cohen-effect-size-h-arcsin-transformation.htm", 2013. [Accessed 10-01-2019].

[41] Robert I. Kabacoff, Ph.D. Power analysis. "https://www.statmethods.net/stats/power.html". [Accessed 15-03-2019].

[42] Ronald L. Wasserstein, Nicole A. Lazar. The asa's statement on p-values: Context, process, and purpose. *The American Statistician*, 70:129–133, 2017.

[43] Helios De Rosario. *pwr: Basic Functions for Power Analysis*. RQDA, 2018. R package version 1.2-2.

[44] Per Runeson and Martin Höst. Guidelines for conducting and reporting case study research in software engineering. *Empirical software engineering*, 14(2):131, 2009.

[45] Russ Winer, Ravi Dhar. *Marketing Management*. Pearson, 2015.

[46] Steven L. Scott. Overview of content experiments: Multi-armed bandit experiments. "https://support.google.com/analytics/answer/2844870?hl=en", 2018. [Accessed 30-10-2018].

[47] Tessa V. West Sean P. Lane, Erin P. Hennes. Workshop. In *Ive Got the Power: How Anyone Can Do a Power Analysis of Any Type of Study Using Simulation*, 2016.

[48] Kim Bartel Sheehan. E-mail survey response rates: A review. *Journal of Computer-Mediated Communication*, 6(2), 2001.

[49] Stephanie. Non parametric data and tests (distribution free tests). "https://www.statisticshowto.datasciencecentral.com/parametric-and-non-parametric-data/". [Accessed 15-05-2019].

[50] Stephanie. Confidence level: What is it? "https://www.statisticshowto.datasciencecentral.com/confidence-level/", October 2014. [Accessed 28-02-2019].

[51] Thomas K. Landauer, Prasad V. Prabhu, Martin G. Helander, P. V. Prabhu. *Handbook of Human-Computer Interaction (2nd Edition)*. Elsevier Science & Technology, Amsterdam, 1997.

[52] CR Wilson VanVoorhis and Betsy L Morgan. Understanding power and rules of thumb for determining sample sizes. *Tutorials in quantitative methods for psychology*, 3(2):43–50, 2007.

[53] Wikipedia. Standard deviation. "https://en.wikipedia.org/wiki/Standard_deviation", February 2019. [Accessed 28-02-2019].

[54] Kevin B Wright. Researching internet-based populations: Advantages and disadvantages of online survey research, online questionnaire authoring software packages, and web survey services. *Journal of computer-mediated communication*, 10(3):JCMC1034, 2005.

[55] Marvin Zelen. The randomization and stratification of patients to clinical trials. *Journal of chronic diseases*, 27(7):365–375, 1974.