Challenges in Applying Continuous Experimentation
A Practitioners' Perspective

Anderson, Kevin; Visser, Denise; Mannen, Jan-Willem; Jiang, Yuxiang; van Deursen, Arie

DOI
10.1109/ICSE-SEIP55303.2022.9793934

Publication date
2022

Document Version
Final published version

Published in
Proceedings - 2022 ACM/IEEE 44th International Conference on Software Engineering

Citation (APA)
Anderson, K., Visser, D., Mannen, J.-W., Jiang, Y., & van Deursen, A. (2022). Challenges in Applying Continuous Experimentation: A Practitioners' Perspective. In *Proceedings - 2022 ACM/IEEE 44th International Conference on Software Engineering: Software Engineering in Practice, ICSE-SEIP 2022* (pp. 107-114). Article 9793934 (Proceedings - International Conference on Software Engineering). IEEE. https://doi.org/10.1109/ICSE-SEIP55303.2022.9793934

Challenges in Applying Continuous Experimentation: A Practitioners' Perspective

Kevin Anderson*
Delft University of Technology / Vista
Utrecht, Netherlands
k.s.anderson@tudelft.nl

Denise Visser
Bol.com
Utrecht, Netherlands
dvisser@bol.com

Jan-Willem Mannen
ING
Amsterdam, Netherlands
jan-willem.mannen@ing.com

Yuxiang Jiang†
Delft University of Technology
Delft, Netherlands
y.jiang-12@student.tudelft.nl

Arie van Deursen
Delft University of Technology
Delft, Netherlands
arie.vandeursen@tudelft.nl

ABSTRACT

Background: Applying Continuous Experimentation on a large scale is not easily achieved. Although the evolution within large tech organisations is well understood, we still lack a good understanding of how to transition a company towards applying more experiments.
Objective: This study investigates how practitioners define, value and apply experimentation, the blockers they experience and what to do to solve these.
Method: We interviewed and surveyed over one hundred practitioners with regards to experimentation perspectives, from a large financial services and e-commerce organization, based in the Netherlands.
Results: Many practitioners have different perspectives on experimentation. The value is well understood. We have learned that the practitioners are blocked by a lack of priority, experience and well functioning tooling. Challenges also arise around dependencies between teams and evaluating experiments with the correct metrics.
Conclusions: Organisation leaders need to start asking for experiment results and investing in infrastructure and processes to actually enable teams to execute experiments and show the value of their work in terms of value for customers and business.

CCS CONCEPTS

• **General and reference** → *Surveys and overviews.*

KEYWORDS

Continuous experimentation, Online controlled experiments, A/B testing, Empirical software engineering, ING, bol.com

*Work completed while at ING.
†Work completed during an internship at ING.

ICSE-SEIP '22, May 21–29, 2022, Pittsburgh, PA, USA

ACM ISBN 978-1-4503-9226-6/22/05.
https://doi.org/10.1145/3510457.3513052

ACM Reference Format:
Kevin Anderson, Denise Visser, Jan-Willem Mannen, Yuxiang Jiang, and Arie van Deursen. 2022. Challenges in Applying Continuous Experimentation: A Practitioners' Perspective. In *44nd International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP '22), May 21–29, 2022, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3510457.3513052

1 INTRODUCTION

Organizations like Microsoft, Google, Facebook and Booking.com use online controlled experiments (A/B tests) to assess the impact of changes made to software products and services [6]. The evolution towards doing continuous experimentation is well understood and documented within the context of these large tech companies [2–4]. Many of these tech companies invest in their own experimentation platform, because it is perceived as critical to their business [7].

This study is performed in the context of a large financial services (ING) and an ecommerce (bol.com) organization. Although both companies are active in different industries, they do share some other characteristics. Both organizations have a strong presence in the Netherlands with well known brands [9]. Both companies have also invested in the creation of in-house built experimentation infrastructure which has been tightly integrated into core systems. In this study we look beyond the technical software engineering challenges around continuous experimentation. Cultural factors are equally important in successful adoption [6].

We expect to add to the increasing academic literature on this topic from the perspective of organisations where continuous experimentation is *not yet* the standard. Especially from the perspective of practitioners who are (or could be) executing experiments.

The remainder of this paper is organized as follows. In Section 2 related work is described. In Section 3 we outline the study design. The results of the case study are described in Section 4. We discuss the results in Section 5 and the threats to validity in Section 6, after which we conclude our paper in Section 7.

2 BACKGROUND AND RELATED WORK

In a mapping study by Ros and Runeson it was concluded that four types of challenges around experimentation already received research attention: technical, statistical, organizational/management, and business challenges [16].

Gupta et al. further identified challenges from 13 organizations that already apply experimentation on a large scale. Even these organizations have challenges around creating and maintaining a culture of experimentation [6].

Lindgren and Münch researched the state of experimentation in 10 software organizations in Finland [15]. Their results show that the state of experimentation is not yet mature and the researchers concluded that the challenges in moving towards continuous experimentation are not on the technical side, but on the organizational level: culture, slow development speed, and difficulties in measuring customer value correctly.

It is important for product and service developers to continuously learn what customers want [5]. Applying continuous experimentation is a way to do that. Yaman et al. defined continuous experimentation as an experiment-driven development approach where critical assumptions are tested iteratively [19].

The studies performed by Yaman within the context of four software companies from the Nordics, show that the transition to continuous experimentation is a learning process where it is important to understand the perspective from practitioners [18]. The human factor plays an important part in adopting experimentation practises and she concluded that individual people have different perspectives of what experiments are.

Fabijan et al. provided actionable steps to grow and keep a culture of experimentation and A/B testing [2]. They presented the concept of a flywheel where initial value from experimentation leads to more investments, which will lead to more value. They pointed out that the most difficult part is actually getting the first traction of this flywheel effect.

Kohavi provided various guidelines on how to deploy experimentation in industry [11, 12]. While his work provided important insights on how to conduct *better* experiments, they offered less organizational guidelines to support organizations that are on the path towards becoming more data and experiment driven.

3 RESEARCH METHOD

The goal of this study is to understand the specific challenges around experimentation in two organizations from different industries, both active in the Netherlands. Both organizations were early adopters of agile methodologies and have taken similar approaches to enabling experimentation with the support and development of in-house build experimentation tooling. At the time of this study the first, second and third authors were all directly involved in developments around the respective experimentation tools at ING and bol.com. Therefore they have a good understanding of the state of experimentation within the organizations.

We formulated four research questions to set more fine-grained directions for our study and to limit its scope.

***RQ1**: How do practitioners define and apply experimentation?*

***RQ2**: What is the perceived value of experimentation?*

***RQ3**: What are blockers in doing more experiments?*

***RQ4**: What should an organization do to solve these blockers?*

The data in this study is collected in two ways; via interviews and a survey. The detailed research design is described in the following two subsections.

3.1 Recruitment and Participants

We recruited participants (n=34) based on personal knowledge of the organization and by asking early participants for references. We defined the following requirements for our participants to ensure a solid perspective of the challenges at hand:

- Mix of people in technical and business related roles
- Mix of people in individual contributor and leadership roles
- Significant professional experience (a minimum of five years)

Only two people declined the invitation to be interviewed, one due to time constraints and another because the person did not feel qualified to talk about the topic. We expect that the personal connection with the interviewers and the ease of doing a short interview over video call, led to the high acceptance rate. See Table 1 for a list of all participants, their role and experience, and the duration of the interview in minutes.

3.2 Interview Procedure and Analysis

The semi-structured interviews were based on a set list of questions based on experience and discussions between the first and third authors. See the appendix for an overview of the questions. During the interview there was room for follow-up questions based on the responses. All interviews (n=34) were conducted via video conferencing and most of them also recorded (when permission was given). During the interview short notes were taken and afterwards the recorded interviews were transcribed for further analysis. All interviews at ING were conducted in January and February 2021, alternately by the first and third authors. The interviews at bol.com were performed, by the second author, in July and August of the same year.

After performing all (26) interviews at ING the first and third authors, analysed the results and formulated a first set of general themes. After the (8) interviews at bol.com, the outcome was discussed by the first and second authors and this led to a further refinement of the general themes. The combined results are presented in section four.

3.3 Survey Procedure and Analysis

The results from the qualitative interviews have been validated via a survey, which was conducted at ING. A total of 868 persons from 4 different departments (tribes) were selected for participation in this survey. The departments were selected based on the scope of their work: digital transformation and improving digital sales and service. The current adoption of experimentation, measured by number of online controlled experiments, is also the highest in two of these departments. The other two are in the middle and bottom tiers of departments executing experiments.

The first and fourth authors jointly setup the questionnaire (see the appendix) and was pilot tested with the third author. After considering all feedback and implementing minor refinements, the questionnaire was distributed via email by the first author. The survey was held between September 28 and October 6. 73 people completed the survey, leading to a response rate of 8.4%. Although this response rate is lower than external benchmarks [1], it is comparable to response rates from other surveys within ING. More importantly, the survey participants are a good representation across ING's departments and job roles.

Table 1: Profile of interview participants

ID	Org.	Role/Function	Exp. (yrs)	Duration (min.)
P1	ING	Lead Customer Experience	12	40
P2	ING	Manager Innovation	12	35
P3	ING	Manager Optimisation	24	41
P4	ING	Manager Data Science	13	27
P5	ING	Product Owner	16	45
P6	ING	Online Marketer	15	31
P7	ING	Online Marketer	10	22
P8	ING	Online Marketer	21	21
P9	ING	Product Owner	10	46
P10	ING	Product Owner	19	42
P11	ING	Innovation Consultant	18	56
P12	ING	Lead Customer Experience	24	23
P13	ING	Online Marketer	8	24
P14	ING	Innovation Consultant	25	46
P15	ING	Director Retail	25	23
P16	ING	Product Owner	14	38
P17	ING	Manager Social Media	6	22
P18	ING	Manager UX	17	36
P19	ING	Director Digital	23	29
P20	ING	Online Marketer	13	47
P21	ING	Sr. Manager Digital	21	23
P22	ING	Product Owner	5	50
P23	ING	Full Stack Engineer	12	35
P24	ING	Sr. iOS Developer	15	39
P25	ING	Full Stack Developer	17	50
P26	ING	Android Developer	15	31
P27	bol.com	Product Manager	15	37
P28	bol.com	Product Manager	10	47
P29	bol.com	Director IT	30	45
P30	bol.com	Manager Customer Service	18	36
P31	bol.com	Test Engineer	22	24
P32	bol.com	Director Business Models	30	25
P33	bol.com	Team Lead Analytics	15	38
P34	bol.com	Software Engineer	5	45

4 RESULTS

This section presents our findings from the interviews and survey.

4.1 Definition and application of experimentation (RQ1)

4.1.1 Definition of experimentation. Within ING an experiment is defined as '*a test observing how customers react in order to validate (or invalidate) business assumptions*' [8]. The interviewees have a broad range of definitions of experimentation. Ranging from: "*simply trying something*" (P2), "*making a change and see what the effect is*" (P5), to "*running a randomized controlled trial, a pilot is not an experiment to me*" (P16). The latter definition is a much more strict definition.

Table 2 shows the results from the survey participants on the multiple-choice question (see Q3 in the Appendix). The standard definition within ING is also the most mentioned concept by the survey participants (88%). To our surprise, 'asking colleagues what they think about an idea', is also seen as part of experimentation by 25 (34%) participants. From the survey we learned that over 60% of respondents already took some form of experimentation training (called 'PACE Academy'). This percentage is roughly the same (68%) in the group that finds "asking colleagues' part of experimentation. This might be due to time passed since attending training or conflicting perceptions.

Table 2: Definition & application of experimentation at ING

Experimentation concept	Partic. #, %
Observing customers reaction to (in)validate assumptions	64 (88%)
Quickly test something before committing to building it completely	58 (79%)
A/B testing	58 (79%)
Trying out something new	56 (77%)
Learning what works	56 (77%)
Developing a hypothesis	45 (62%)
Interviewing the target audience	38 (52%)
Incremental improvements	37 (51%)
Working with the PACE canvases	28 (38%)
Asking colleagues what they think about an idea	25 (34%)
Changing a webpage on ING.be/nl	15 (21%)
De-risking a project	13 (18%)

For now, we can conclude that *experimentation* means, next to the default definition, many other things to different people.

4.1.2 Application of experimentation. All participants from the interviews mention concrete examples of experiments that they or their team have executed. This ranges from A/B tests, pilot groups to deploying data science models with control groups in place.

From the broader survey we learn that many respondents (36%) have not executed an experiment in the last 6 months. 33% have executed an experiment in the last month or sprint. See Table 3.

4.2 Perceived value of experimentation (RQ2)

From the interviews we captured about four distinct categories where experimentation brings value to the participant: focus on value, risk mitigation, team alignment and intrinsic benefits on a personal level.

4.2.1 Focus on value. Colleagues are taught that "*experiments help to reduce the uncertainty in our backlog, ensuring our scarce resources are only used for the things that really matter for our customers*" [8]. Especially knowing what kind of value you deliver to customers, is something that is often mentioned by participants. Participant P4 said it like this: "*... knowing that you are bringing value to the customer and knowing what that value is*". And participant P33 described it like this: "*Understand which levers you can pull to add value for the customer and the company*". Other participants stressed the value of validating assumptions and moving away from opinions: "*Nobody has a monopoly on wisdom, we need to validate the*

assumptions we have. That means testing different versions and learn from that. It's a continuous process of validating assumptions", said participant P22. Similar remarks are made by participant P19: "*I believe in continuous improvement. We need to move away from thinking in opinions and thinking more in facts. Everything is constantly changing, so we have to get this way of thinking in the genes*". Participant P21 immediately shared a clear example of where the application of experimentation led to significant results: "*Basically through continuously doing experiments and monitoring we have been able to improve process X from 45% in 2019 to 70% now*". The focus on value delivered that experimentation brings, is clearly a big benefit perceived by the participants.

4.2.2 Risk mitigation. Another reason to use experimentation is to de-risk a project. Like participant P12 says: "*I see experimentation as a way to make sure you validate the riskiest assumptions with customers in an early stage of any project"*. This can prevent unexpected behavior in later stages or the change to mitigate the identified risk. To our surprise, the term 'de-risking a project' resonated the least with the survey participants (see table 2). Only 13 (18%) think about this when they are asked to define experimentation.

4.2.3 Team alignment. Experiments can give objective information about the impact and direction of certain effects. Some participants emphasize the value of having this objective data in the context of discussions within, and between teams. Participant P5 said: "*we simply have better discussions*". And participant P34 stresses that "*experiments give clear answers to specific questions and discussions we are having with other teams.*" This can enormously reduce the number of opinion based discussions people in and between teams have.

4.2.4 Intrinsic benefits. The fourth benefit that came out of our round of interviews is more on a personal level. Some participants focus on the intrinsic value that doing experimentation brings to them. Like participant P5 says: "*I simply enjoy to see if something works as intended. And my life as Product Owner becomes easier, because everyone has an opinion, but with data from experiments we simply have better discussions"*. And according to participant P6: "*Experiments give me certainty, certainty about what is happening. I am creative, open minded and I like facts. It gives me strength when I know what the effect is. I don't like to bluff, I like to see the facts.*" Curiosity is also a term that was mentioned often: "*I'm curious to find out if people understand what we came up with.*", said participant P18. Finally, participant P32 didn't see any intrinsic benefits: "*the benefits are for the company, not for me personally"*.

4.3 Experimentation blockers (RQ3)

With so many strong and clear benefits, the next question focused on the blockers of doing large scale experimentation.

4.3.1 No priority. The most common reason that was mentioned was simply not having enough priority to create experiments. Participant P1 summarizes it well: "*Focus is often on delivery, experiments or other ways of measuring impact does not get sufficient priority*". And participant P30 pinpoints that: "*Often, the solution to a problem has already been devised. Then it's about prioritizing solutions, instead of problems to work on"*. Next to focus on delivery and prioritizing solutions, sometimes priorities are temporarily shifted. Participant P25 explains: "*Focus for our teams is to maintain the feature until an important migration is done. There is no room for improvements now"*.

Table 3: Have you or your squad executed any experiment during the last 6 months?

Frequency	Partic. #, %
No experiment in last 6 months	26 (36%)
Yes, once per month	17 (23%)
Yes, once per quarter	14 (19%)
Yes, once in the past 6 months	9 (12%)
Yes, at least once per sprint	7 (10%)

4.3.2 Dependencies. Almost all participants mentioned having too many dependencies as a blocker in executing more experiments. Participant P10 explained: "*For many experiments I am dependent on a Data Analyst, and it takes them a lot of time. So it feels I need to bother someone else with it"*. And participant P6 says: "*I am dependent on IT development resources in other teams to make changes and launch experiments*". Next to mentioned dependencies on Data Analysts and Software Engineers, other participants mentioned dependencies with Legal and User Experience Experts.

4.3.3 No experience. From the survey participants 64% (n=47) say they have received training around experimentation. Also 77% indicated that they have discussed experimentation with their team. But also almost half of the survey respondents indicate they (or their squad) have executed no (36%) or just 1 (12%) experiment in the last 6 months. More teams are discussing than doing experimentation. The lack of experience is not helping in increasing the experiment velocity.

4.3.4 No or hard to use tooling (functionality). Participants point to missing or broken functionality in the current experimentation tooling. Some functionality had been present, but is currently not working properly. Participant P13 even points out that "*At the moment we don't have access to an A/B testing tool, so we are doing one version at the time: first two weeks version A and then two weeks version B. That is of course not a pure way of testing*". This participant clearly has the will to experiment, but is not supported with the correct tooling.

4.3.5 Issues with metrics. Another issue that participants mention is the fact that they experience issues with evaluating their experiments in the correct way. Some type of metrics are not available, which forces people to use less relevant metrics for the evaluation of experiments. Some teams have trouble coming up with proper measurements at all: "*it is often hard to formulate clear measurable goals. This makes it difficult to choose a meaningful metric for evaluation of experiments*" says participant P16. Another participant (P20) says "*I have continuously doubts about the metrics I'm seeing and I need to do the troubleshooting myself*".

4.4 Improvements (RQ4)

In this section we will answer the question: how can we improve the situation around the blockers identified in section 4.2. This is based on the answers from the study participants, but also on our own understanding of the challenges.

4.4.1 Leadership priority. The remarks around a lack of priority are in stark contrast with the identified value of experimentation. It seems that everyone during our conversations was convinced that the organization should be doing large scale experimentation, but somehow they are not doing it (enough).

Participant P28 says it like this: "*It is not expected of us to run experiments. We manage internal expectations based on the 'WHAT', and less on the contribution to KPI's*".

This suggest that leaders should start challenging teams to evaluate developed features based on quantifiable results, instead of only on timely delivery. One Senior Manager (participant 21) acknowledged that "*We need to accept the trade-offs. In the beginning we will probably go slower, but experimentation will eventually make us go faster*". This suggests that currently the teams that have to deal with many dependencies are already satisfied when they have shipped a feature to production.

Setting priorities could also help in forming sufficient self supported teams. This challenge is about speed of development. The organization needs to have as many independent and empowered teams that are sufficiently staffed to live up to their purpose. As participant P19 indicates: "*Pressure on [feature] delivery is high. This limits us to experiment. Also some teams have challenges with getting the output we seek. We are looking into ways to optimise this. This will also make room for more validation*".

Although this blocker is not only an issue for experimentation itself. This is hindering development of products and services in general. In this setting experimentation is being perceived as adding an extra layer of complexity to the product development process. Therefore it is extra important to make running an experiment as easy as possible.

4.4.2 Integrated tool. The second challenge organizations should invest in is in making the tooling to execute an experiment as easy as possible. As participant P19 notes: "*People find it difficult to quickly validate hypotheses. Eventually people just start building. Our tooling is now not really plug-and-play, we lack easy to use tooling*". A seamless integration of experimentation tooling in the software development lifecycle can lower the cost of setting up an experiment [2]. This continuous drive in trying to lower the cost of an experiment needs a dedicated team to (in the first place) enable it and after that, keep it up and running, and continuously improve this system. Participant P1 said it like this: "*All the friction you can take away, you should. People are a bit lazy by nature, you have to make it as easy as possible*".

4.4.3 Trustworthy results. Make sure quality controls are in place and that issues around measuring are solved in a timely fashion. This also needs a dedicated team that continuously monitors the health of the experimentation tool. To prevent each and everyone to solve issues by themselves, like participant (P20) indicated "*I have continuously doubts about the metrics I'm seeing and I need to do the troubleshooting myself*".

4.4.4 Education. There needs to be training on the how to do experimentation and coming to a common understanding of what it exactly is. As the research shows, running an experiment means several things to people. We expect that a stricter definition will make it easier to get understanding of what it actually means to do experimentation. Participant P2 stresses the importance of accessible learning opportunities: "*All PACE modules are now freely available. This training was very expensive, now anyone can join for free whenever they like via our online portal*". Participant P2 also says that training people is a good start: "*It's a combination of know-how, mindset and support from the top. We can start by investing in more training*". Participant P11 does point out that training alone might not be enough: "*Some people will still need assistance in setting up their first experiment*".

4.4.5 Share learnings. By sharing the learnings from experiments across the organisation, we show that gut feeling is often not correct. This can act as a new trigger to start experimenting. Next to that, continuously showing the value from experiments will ensure future investments in the experimentation program, as also shown by Fabijan et al. [2]. As participant P13 says: "*We should organise more sessions where we can share experiences around experiments, what works, what does not work. And what can other teams test as well*". And participant P22 says "*We need a sharing platform to store the learnings from past experiments. This can act as inspiration for others*".

5 DISCUSSION

The outcomes of our interviews and survey as presented in this paper highlight how practitioners *define and apply* experimentation, and what *challenges* they face in applying experimentation more often in their daily work. From this, we suggest several lines of action, divided in recommendations for practitioners (companies and vendors) and recommendations specific to researchers and educators.

5.1 Recommendation to Practitioners

First, practitioners should make sure that everyone within an organization is aware of what experimentation is, and what it is not. For example, asking colleagues what they think about an idea, might provide valuable information, but it lacks the core concept of simply observing behavior, without explicitly asking about thoughts and feelings. The many different meanings experimentation currently seem to have to different people makes it harder to steer the desired behavior around experimentation. We also recommend to include the concept of risk mitigation in applying experimentation. These topics should all be addressed in training.

Second, the transition towards continuous experimentation starts with company senior leaders asking for experiment results. As long as projects are mostly being steered based on delivery timelines [14], there is less incentive for teams to start or increase the rate of experimentation. We recommend leaders to start measuring the impact of teams in terms of outcome metrics, not purely on output. This includes challenging teams on how they have come to decisions by asking 'How do you know?' and 'How do we know we will be right?'.

Third, organizations cannot expect large and sustainable impact from experimentation when their efforts around training and infrastructure are not organised from one dedicated team or centre of excellence. This relates to the blocking issues around priority (Section 4.3.1). Organizations that take experimentation seriously, should invest in this important area. Good examples here are Booking.com [10] and LinkedIn [17]: both companies have multiple dedicated teams in place to build infrastructure, train people and facilitate large scale experimentation.

Fourth, the barriers identified around dependencies (Section 4.3.2) pose impediments to the successful adoption of experimentation. These blocking issues go beyond experimentation. For example, if teams are too dependent on others, this may be an issue to address *first*, before asking teams to start to execute (more) experiments. Resource allocation and team structure go beyond the application of experimentation. For example, research by Kula et al. indicates that managing or reducing task dependencies is a key factor contributing to on time delivery in large-scale agile development [13]. The blockers identified in this study stress the importance of this topic for leaders to successfully deal with.

5.2 Recommendation to Researchers and Educators

We advise researchers to team up with organizations on the path towards continuous experimentation. It can lead to a better understanding of the sometimes messy practice many teams operate in. More research is needed to better understand what interventions are most effective in getting teams to adopt continuous experimentation.

We also encourage cross industry collaboration between researchers and companies from different industries. This way we can better learn which challenges are organization specific and which are common across industries.

Lastly, academic educators might want to consider experimentation a first class citizen in undergraduate software engineering courses. This calls for lab work in which feature delivery is not just based on time taken and code quality, but on actually conducted experiments demonstrating that feature A leads to a better end user experience than feature B. This tight connection between coding and end user satisfaction may be hard to achieve as it requires the involvement of end users in an academic course, but our findings suggest that such an experience would prepare students well for their next position in industry.

6 THREATS TO VALIDITY

In this section we discuss the possible threats to validity and the action we took to mitigate them.

6.1 Threats to Internal Validity

Possible threats to internal validity mostly come from the interview process we propose and the analysis of the semi-structured interviews performed. We mitigated this by sufficient test rounds within the author group, but also with the participants. For example, the first four interviews were jointly attended by the first and third author. Both authors took turns in interviewing and observing. Afterwards they jointly reflected on the process. This way we ensured a similar approach in interviewing.

We tried to recruit the right people and not only people with a positive outlook on experimentation. To avoid selection bias, the term 'experimentation' was not mentioned in the survey invitation email. However, since it was framed as "delivering differentiating customer experiences", this could also have led to attract specific colleagues. This might have influenced the number of respondents. As mentioned earlier, the participants are a good representation of departments and job roles.

6.2 Threats to External Validity

Naturally, our results come from a limited set of organizations in a specific region. Likely when we expand with companies of different sizes, from different application domains, or from other regions we will find additional challenges around adopting continuous experimentation, as well as new challenges that the people in the two organisations in our study have not yet been facing. Our results serve as a starting point to conduct such further studies.

7 CONCLUSIONS

Applying continuous experimentation on a large scale is not easily achieved. This paper describes the perspectives and challenges in a large financial services and an e-commerce organisation, after interviewing 34 practitioners from these organisations. Next to that we surveyed 73 practitioners to corroborate the results from the interviews.

Our results point out that many practitioners have different perspectives on what experimentation exactly is. The potential value is understood: focus on customer value, aligning teams, next to some intrinsic benefits. Although risk mitigation was often mentioned during the interviews as a potential benefit, only 18% (22) of the survey participants agreed with this. Unfortunately, experimentation is not being applied often. Our findings indicate that practitioners are blocked by a lack of priority, experience and/or well functioning and easy-to-use tooling, and by too many dependencies between teams as well as the choice of metrics in evaluating experiments.

The study emphasizes the need for a common understanding of what experimentation is. Leadership needs to start asking about experiment results. This requires investing in infrastructure and processes to actually enable teams to execute experiments that show the value of their work in terms of value for customers and businesses.

We sincerely hope that this paper can spark more research into organizations on their path to applying continuous experimentation on a large scale. We believe this study can be an important next step in helping organisations on this journey.

ACKNOWLEDGMENTS

The authors would like to thank all participants for their willing contributions to this project. We also wish to thank the three reviewers and Lukas Vermeer for their constructive feedback. This paper was partially funded by ICAI AI for Fintech Research.

REFERENCES

[1] Yehuda Baruch and Brooks C. Holtom. 2008. Survey response rate levels and trends in organizational research. *Human relations* 61, 8 (2008), 1139–1160. Publisher: Sage Publications Sage UK: London, England.

[2] Aleksander Fabijan, Benjamin Arai, Pavel Dmitriev, and Lukas Vermeer. 2021. *It takes a Flywheel to Fly: Kickstarting and Growing the A/B testing Momentum at Scale.*

[3] Aleksander Fabijan, Pavel Dmitriev, Helena Holmström Olsson, and Jan Bosch. 2017. The evolution of continuous experimentation in software product development: from data to a data-driven organization at scale. In *Proceedings of the 39th International Conference on Software Engineering*. IEEE Press, 770–780.

[4] Fabian Fagerholm, Alejandro Sanchez Guinea, Hanna Mäenpää, and Jürgen Münch. 2014. Building blocks for continuous experimentation. In *Proceedings of the 1st International Workshop on Rapid Continuous Software Engineering (RCoSE 2014)*. Association for Computing Machinery, Hyderabad, India, 26–35. https://doi.org/10.1145/2593812.2593816

[5] Fabian Fagerholm, Alejandro Sanchez Guinea, Hanna Mäenpää, and Jürgen Münch. 2017. The RIGHT model for continuous experimentation. *Journal of Systems and Software* 123 (2017), 292–305. Publisher: Elsevier.

[6] Somit Gupta, Ronny Kohavi, Diane Tang, Ya Xu, Reid Andersen, Eytan Bakshy, Niall Cardin, Sumita Chandran, Nanyu Chen, Dominic Coey, and others. 2019. Top challenges from the first practical online controlled experiments summit. *ACM SIGKDD Explorations Newsletter* 21, 1 (2019), 20–35. Publisher: ACM New York, NY, USA.

[7] Somit Gupta, Lucy Ulanova, Sumit Bhardwaj, Pavel Dmitriev, Paul Raff, and Aleksander Fabijan. 2018. The Anatomy of a Large-Scale Experimentation Platform. In *2018 IEEE International Conference on Software Architecture (ICSA)*. IEEE, 1–109.

[8] ING. 2020. Experiment Design Playbook. internal document.

[9] Kantar. 2020. *BrandZ Top 30 Most Valuable Dutch Brands 2021*. Technical Report. Kantar. https://www.rankingthebrands.com/PDF/Brandz%20Most%20Valuable%20Netherlands%20Brands%202021%20top%2030.pdf

[10] Raphael Lopez Kaufman, Jegar Pitchforth, and Lukas Vermeer. 2017. Democratizing online controlled experiments at Booking. com. *arXiv preprint arXiv:1710.08217* (2017).

[11] Ron Kohavi, Alex Deng, Roger Longbotham, and Ya Xu. 2014. Seven rules of thumb for web site experimenters. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '14)*. Association for Computing Machinery, New York, New York, USA, 1857–1866. https://doi.org/10.1145/2623330.2623341

[12] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M. Henne. 2009. Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery* 18, 1 (Feb. 2009), 140–181. https://doi.org/10.1007/s10618-008-0114-1

[13] Elvan Kula, Eric Greuter, Arie Van Deursen, and Gousios Georgios. 2021. Factors Affecting On-Time Delivery in Large-Scale Agile Software Development. *IEEE Transactions on Software Engineering* (2021), 1–1. https://doi.org/10.1109/TSE.2021.3101192

[14] Elvan Kula, Ayushi Rastogi, Hennie Huijgens, Arie van Deursen, and Georgios Gousios. 2019. Releasing fast and slow: an exploratory case study at ING. In *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. ACM, Tallinn Estonia, 785–795. https://doi.org/10.1145/3338906.3338978

[15] Eveliina Lindgren and Jürgen Münch. 2016. Raising the odds of success: the current state of experimentation in product development. *Information and Software Technology* 77 (Sept. 2016), 80–91. https://doi.org/10.1016/j.infsof.2016.04.008

[16] Rasmus Ros and Per Runeson. 2018. Continuous experimentation and a/b testing: A mapping study. In *2018 IEEE/ACM 4th International Workshop on Rapid Continuous Software Engineering (RCoSE)*. IEEE, 35–41.

[17] Diane Tang, Ashish Agarwal, Deirdre O'Brien, and Mike Meyer. 2010. Overlapping experiment infrastructure: More, better, faster experimentation. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 17–26.

[18] Sezin Gizem Yaman. 2019. Initiating the Transition towards Continuous Experimentation : Empirical Studies with Software Development Teams and Practitioners. (Oct. 2019). https://helda.helsinki.fi/handle/10138/305855 Accepted: 2019-10-04T06:00:19Z ISBN: 9789515155436 Publisher: Helsingin yliopisto.

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

Interview questions

Q1. *Questions on personal level*

1. Who are you? And what is your role at the company? Can you briefly explain what you do?
2. When we talk about experimentation, what is it according to you? How would you define it?
3. What is your own experience with doing experiments?
4. How and when do you apply it in your daily work?
5. What is the value/advantages of experimentation to you?
6. What is blocking you to do more experiments?

Q2. *Questions on organizational level*

1. What is the experience with doing experiments in your department
2. How and when do people apply it in their daily work?
3. What is the value/advantages of experimentation to the people in your department?
4. What is blocking people to do more experiments?
5. How do we improve on experimentation?
6. Who else should we interview for this project?

Survey questions

Q3. *What is experimentation according to you?*

1. Quickly test something before committing to building it completely
2. Observing customers reaction to (in)validate assumptions
3. Asking colleagues what they think about an idea
4. Changing a webpage on ING.be/nl
5. Interviewing the target audience
6. Working with the PACE canvases
7. Incremental improvements
8. Trying out something new
9. Developing a hypothesis
10. Learning what works
11. De-risking a project
12. A/B testing
13. Other

Q4. *Have you or your squad discussed experimentation during the last 6 months? [Yes/No]*

Q5. *Have you taken any training for experimentation? [Yes/no]*

Q6. *Have you or your squad executed any experiment during the last 6 months?*

1. Yes, at least once per sprint
2. Yes, once per month
3. Yes, once per quarter
4. Yes, once in the past 6 months
5. No

Q7. *To what extent do you agree with the following statements? [Strongly disagree, Disagree, Neither agree nor disagree, Agree, Strongly agree, N/A]*

1. The current state of experimentation at my squad/tribe needs to be improved
2. Experimentation means more unnecessary workload for my job. Experimentation takes up too much time, while decisions need to be made quickly.
3. Agile way of working asking us to move fast, which leaves us no time for experimentation.
4. I'm confident enough to perform experiments.
5. Failure and invalidation of my ideas/assumptions are not encouraged by my KPIs or my performance appraisal.
6. My leadership does not like it when I fail or my ideas/assumptions get invalidated.
7. I feel very frustrated when my idea or assumption is proved to be wrong.
8. I feel the need to always be right at my job.
9. I have the skills and knowledge to execute an experiment.
10. I know where to find support or tooling to execute an experiment.
11. I (and/or my squad) have been given enough time, budget, and priority to conduct experiments.
12. We as a squad talked about experimentation with other squads/tribes we collaborated with.
13. Migration projects and requests from other squads/tribes acted as an obstacle for us to make changes to our way of working.
14. Experimentation is someone else's job in my squad.
15. My role is not involved in experimentation.
16. I'm not sure where or how to start with experimentation.
17. My squad is a DevOps squad that takes and executes external requests, there is no room for us to experiment.
18. I find the tooling available within ING for experimentation difficult to use. There are other more important objectives at ING like cost saving or fast delivery. Therefore experimentation has to take a back seat.
19. I only have internal customers, so experimentation is not for me.
20. Experiment (loop) is too big to fit into our daily work.
21. My scope of work does not allow me to conduct experiments. Experimentation is only for big "innovation" projects that happen.
22. I don't see how experimentation can help me do my job better.
23. I find it difficult to connect experimentation with processes at ING.
24. I tried to convince my squad/tribe leadership in order to execute more experiments.
25. My Tribe leadership frequently asked for evidence for decisions concerning products and marketing.
26. I know what customers want based on my working experience.
27. Launching new feature/releasing new campaign is more appreciated than optimizing existing ones within ING.
28. I know the vision of my squad and/or tribe for experimentation.
29. My initiative of increasing experimentation is supported by my squad and tribe.

Q8. *What are the barriers for you (your squad) in conducting experimentation? [open question]*

Q9. *Do you have any other thoughts or remarks? [open question]*

Q10. *Which Tribe are you part of?*

Q11. *Are you based in Belgium or the Netherlands?*

Q12. *What is your role?*