The Journal of Systems and Software 211 (2024) 112011

Contents lists available at ScienceDirect

The Journal of Systems & Software

journal homepage: www.elsevier.com/locate/jss

A/B testing: A systematic literature review ☆

Federico Quin [a], Danny Weyns [b,a,*], Matthias Galster [c], Camila Costa Silva [c]

[a] Distrinet, KU Leuven, Celestijnenlaan 200A, Leuven, 3000, Belgium
[b] Linnaeus University, Universitetsplatsen 1, Växjö, 352 52, Sweden
[c] University of Canterbury, 69 Creyke Road, Christchurch, 8140, New Zealand

ARTICLE INFO

Keywords:
A/B testing
Systematic literature review
A/B test engineering

ABSTRACT

A/B testing, also referred to as online controlled experimentation or continuous experimentation, is a form of hypothesis testing where two variants of a piece of software are compared in the field from an end user's point of view. A/B testing is widely used in practice to enable data-driven decision making for software development. While a few studies have explored different facets of research on A/B testing, no comprehensive study has been conducted on the state-of-the-art in A/B testing. Such a study is crucial to provide a systematic overview of the field of A/B testing driving future research forward. To address this gap and provide an overview of the state-of-the-art in A/B testing, this paper reports the results of a systematic literature review that analyzed primary studies. The research questions focused on the subject of A/B testing, how A/B tests are designed and executed, what roles stakeholders have in this process, and the open challenges in the area. Analysis of the extracted data shows that the main targets of A/B testing are algorithms, visual elements, and workflow and processes. Single classic A/B tests are the dominating type of tests, primarily based in hypothesis tests. Stakeholders have three main roles in the design of A/B tests: concept designer, experiment architect, and setup technician. The primary types of data collected during the execution of A/B tests are product/system data, user-centric data, and spatio-temporal data. The dominating use of the test results are feature selection, feature rollout, continued feature development, and subsequent A/B test design. Stakeholders have two main roles during A/B test execution: experiment coordinator and experiment assessor. The main reported open problems are related to the enhancement of proposed approaches and their usability. From our study we derived three interesting lines for future research: strengthen the adoption of statistical methods in A/B testing, improving the process of A/B testing, and enhancing the automation of A/B testing.

1. Introduction

Iterative software development and time to market are crucial to the success of software companies. Central to this is innovation by exploring new software features or experimenting with software changes. In order to enable such innovation in practice, software companies often employ A/B testing (Kohavi and Longbotham, 2017; Fabijan et al., 2017b; Li et al., 2019; Goswami et al., 2015). A/B testing, also referred to as online controlled experimentation or continuous experimentation, is a form of hypothesis testing where two variants of a piece of software are evaluated in the field (ranging from variants with a slightly altered GUI layout to variants of software with new features). In particular, the merit of the two variants are analyzed using metrics such as click rates of visitors of websites, members' lifetime values (LTV) in a subscription service, and user conversions in marketing (Johari et al., 2017; Wang et al., 2019; Duan et al., 2021). A/B testing is extensively used in practice, including large and popular tech companies such as Google, Meta, LinkedIn, and Microsoft (Ha-Thuc et al., 2020; Wang et al., 2019; Li et al., 2015; Xu and Chen, 2016; Kohavi and Thomke, 2017; Gupta et al., 2019).

Even though A/B testing is commonly used in practice, to the best of our knowledge, no comprehensive empirically grounded study has been conducted on the state-of-the-art (i.e., state-of-research in contrast to state-of-the-practice) in A/B testing. Such a study is crucial to provide a systematic overview of the field of A/B testing to drive future research forward. Three earlier studies (Auer et al., 2021; Auer and Felderer, 2018; Ros and Runeson, 2018) explored a number of aspects of research on A/B testing, such as research topics, type of experiments in A/B testing, and A/B tooling and metrics. Yet, these studies do not provide

☆ Editor: Antonia Bertolino.

* Corresponding author.
E-mail addresses: federico-quin@hotmail.com (F. Quin), danny.weyns@lnu.se (D. Weyns), matthias.galster@canterbury.ac.nz (M. Galster), camila.costasilva@canterbury.ac.nz (C.C. Silva).

https://doi.org/10.1016/j.jss.2024.112011
Received 8 August 2023; Received in revised form 15 January 2024; Accepted 20 February 2024
Available online 22 February 2024

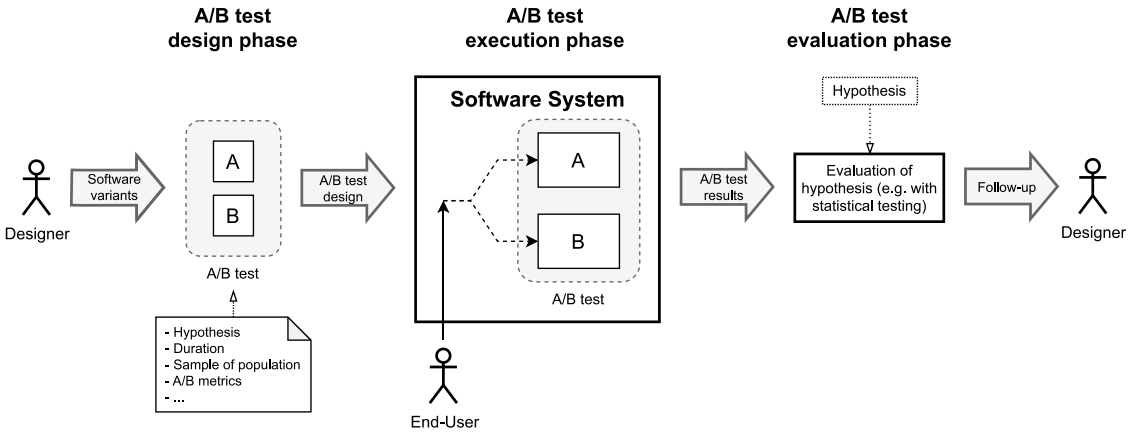

Fig. 1. General A/B testing process.

a comprehensive overview of the state-of-the-art that provides deeper insights in the types of targets to which A/B testing is applied, the roles of stakeholders in the design of A/B tests, the execution of the tests, and the usage of the test results. These insights are key to position and understand A/B testing in the broader picture of software engineering. To tackle this issue, we performed a systematic literature review (Keele et al., 2007). Our study aims to provide insights on the state of research in A/B testing as a basis to guide future research. Practitioners may also benefit from the study to identify potential improvements of A/B testing in their daily practices.

The concrete contributions of this systematic literature review are:

- Insights into the targets and application domains in which A/B testing is conducted.
- Insights into the design of A/B tests, and the roles stakeholders take in designing the tests.
- Insights into aspects of executing A/B tests, and the roles stakeholders take in this process.
- Consolidation of open problems and opportunities for future research directions in A/B testing.

The remainder of this paper is structured as follows. Section 2 provides a brief introduction to A/B testing and discusses related secondary studies. In Section 3, we outline the research questions and summarize the methodology we used. Section 4 then presents the results, providing an answer to each research question. In Section 5, we reflect on the results of the study, report insights, outline opportunities for future research, and outline threats to validity. Finally, Section 6 concludes the paper.

2. Background and related work

2.1. Background

A/B testing is a method where two software variants, denoted as variant A and variant B, are compared by evaluating the merit of the variants through exposure to the end-users of the system (Siroker and Koomen, 2013). To compare the variants, a hypothesis is formulated together with an experiment to test it, i.e., the actual A/B test. As opposed to regular software testing, A/B testing takes place in live systems. Fig. 1 shows the general process of A/B testing with three main phases.

The first phase of A/B testing concerns the design of an A/B test. In this experiment design, a range of parameters is specified, such as: the hypothesis, the sample of the population the experiment should be targeted to, the duration of the experiment, and the A/B metrics that are collected during the experiment. The A/B metrics are used to determine the merit of each variant during the experiment. Examples of A/B metrics include the click-through rate (CTR), number of clicks, and number of sessions (Drutsa et al., 2015b).

The second phase of A/B testing consists of the execution of the A/B test in the running software system. Both variants are deployed in a live system, and the sample of the population is split among both variants. During the execution, the system keeps track of relevant data to evaluate the experiment after it finishes (according to the specified duration). Relevant data may directly correspond to the specified A/B metrics, or it may indirectly enable advanced analysis in the evaluation stage to gain additional insights from the conducted A/B tests.

The third phase of A/B testing comprises the evaluation of the experiment. After the A/B test is finished, the original hypothesis is tested, typically with a statistical test, such as a students test or Welsh's t-test (Gupta and Chokshi, 2020; Vanderdonckt et al., 2019). Based on the outcome of the test, the designer can then take follow-up actions, for instance initiating a rollout of a feature to the entire population or designing new A/B variants to test in subsequent A/B tests.

2.1.1. Controlled experiments vs A/B testing

Traditionally, a controlled experiment is an empirical method that enables to systematically test a hypothesis (Creswell and Guetterman, 2018). Two types of variables are distinguished in controlled experiments: independent and dependent variables. Independent variables are variables that are controlled during the experiment to test the hypothesis, for instance, a state of the art and a newly proposed approach to solve a particular design problem by a control group and a treatment group respectively. Dependent variables are variables that are measured during the experiment to compare the results of both the control and treatment group, for instance, the fault density and productivity obtained in a design task. After conducting the experiment, the hypothesis is tested and conclusions are drawn based on the results; for instance, a newly proposed design approach has a significantly lower fault density compared to the state of the art approach, but more research is required concerning the productivity. Controlled experimentation is widely used across all types of scientific fields, such as psychology (Costa et al., 2011), pharmaceutics (Masuda et al., 2022), education (Creswell and Guetterman, 2018), and nowadays also in software engineering (Siegmund et al., 2015; Daneva et al., 2014; Galster and Weyns, 2016).

Whereas controlled experiments are typically performed *offline in a controlled setting*, A/B testing uses controlled experiments to evaluate software features or variants on the *end-users of a running system*. For this reason, A/B testing is often referred to as online controlled experimentation (Kohavi et al., 2020; Fabijan et al., 2020). The aim of A/B testing lies in testing hypotheses in live software systems where end-users of these systems form the participants or population of the experiment. Examples of hypotheses that are tested in A/B testing often relate to improving user experience (UX) (Renz et al., 2016), improving

user interface (UI) design (Walper et al., 2020), improving user click rates (Aharon et al., 2019b), or evaluating non functional requirements in distributed services (Bakshy and Frachtenberg, 2015).

2.1.2. DevOps and A/B testing

Development Operations (DevOps in short) has gained popularity in recent years (de Feijter et al., 2018). DevOps consists of a set of practices, tools, and guidelines to efficiently and effectively manage and carry out different tasks during software life-cycles. This ranges from the process of software development to the deployment and management of software at runtime. Automation of software processes plays a central part of DevOps to make life easier for developers and ease the burden of software development in general.

Common practices that are part of the DevOps lexicon are continuous integration and continuous deployment (CICD in short) (Humble and Farley, 2010). CICD consists of the automation of software testing, software integration and building, and deployment of software, effectively reducing manual labor required by developers and easing the burden of deploying software. In a similar vein, continuous experimentation (Yaman et al., 2017) aims at continuously setting up experiments in software systems to test new software variants. Put differently, continuous experimentation enriches the software development process by enabling a data-driven development approach (e.g., by measuring user satisfaction of new software features early on in development). To achieve this, A/B testing is used to setup and evaluate online controlled experiments in the software system. Fabijan et al. (2017b) for example perform a case study on the evolution of scaling up continuous experimentation at Microsoft, providing guidelines for other companies to conduct continuous experimentation.

2.2. Related secondary studies

We start with a summary of secondary studies related to the literature review presented in this paper. Then we pinpoint the aim of the study presented in this paper to provide a systematic overview of the state-of-the-art in A/B testing.

2.2.1. Summary of related reviews

We grouped related studies into three classes: studies with a focus on technical aspects of A/B testing, studies focusing on social aspects of A/B testing, and studies concerned with A/B testing in specific domains.

Technical aspects of A/B testing. Rodríguez et al. (2017) performed a systematic mapping study on continuous deployment of software intensive services and products. The authors identify continuous and rapid experimentation as one of the factors that characterize continuous deployment, and elaborate on this through the lens of the deployment of these experiments and DevOps practices associated with it. Ros and Runeson (2018) put forward a mapping study on continuous experimentation and A/B testing. The authors explore research topics, organizations that employ A/B testing, and take a deeper look at the type of experimentation that is conducted. Auer and Felderer (2018) conducted a systematic mapping study on continuous experimentation. The authors put a focus on the research topics, contributions, and research types, collaboration between industry and academia, trends in publications, popularity in publications on A/B testing, venues, and paper citations. Recently, Auer et al. (2021) presented a systematic literature review on A/B testing and continuous experimentation, leveraging the results from previous mapping studies (Ros and Runeson, 2018; Auer and Felderer, 2018). The authors apply forward snowballing on a set of papers to compose the list of primary studies for the review. They then explore the core constituents of a continuous experimentation framework, and the challenges and benefits of continuous experimentation. Closely related, Erthal et al. (2022) conducted a literature review by applying an ad-hoc search, followed by snowballing on the initial set of identified papers. The study places emphasis on defining continuous experimentation and exploring its associated processes. While the authors acknowledge A/B testing as one of the strategies for achieving continuous experimentation, this literature review does not delve into the technical aspects of A/B testing.

Social aspects of A/B testing. An important social aspect of A/B testing is obtaining user feedback. A significant portion of A/B tests revolves around prioritizing and optimizing the user experience. We identified two studies that focus on this social aspect. Fabijan et al. (2015) present a literature review on customer feedback and data collection techniques in the context of software research and development. The authors highlight existing techniques in the literature to obtain customer feedback and organize data collection, in which software development stages the techniques are used, and what the main challenges and limitations are for the techniques. One of the techniques outlined by the authors is A/B testing, which can serve as an valuable tool to obtain user feedback on prototypes. Fabijan et al. (2016) discuss challenges and implications of the lack of sharing customer data within large organizations. One specific case presented by the authors underpins critical issues that manifest from not sharing qualitative customer feedback in the pre-development stage with the development stage, forcing developers to repeat the collection of user feedback or to develop products without this information.

A/B testing in specific domains. Beyond A/B testing at Internet-based companies, the use of A/B testing is reported in various other domains. An example is the domain of embedded systems. Mattos et al. (2018) explore challenges and strategies for continuous experimentation in embedded systems, providing both industrial- and research perspectives. Another domain is Cyber-Physical Systems (CPS). Giaimo et al. (2020) present a systematic literature review on the state-of-the-art of continuous experimentation in CPS, concluding that the literature focuses more on presented challenges rather than proposing solutions to the challenges.

Summary. Existing secondary studies examined A/B testing with a focus on realizing tests, associated processes, and the types of experimentation conducted. However, these studies have a particular focus, or they lack a rigorous search process to identify relevant studies. Existing studies fall short in providing insights in the *target* of A/B testing (i.e., "what" is the subject of testing), the *roles of stakeholders* in designing and executing A/B tests, and the *utilization* of A/B test results.

2.2.2. Aim of the study

To tackle the limitations of existing studies, we performed an in-depth literature study. We define the aim of this study using the Goal Question Metric (GQM) approach (Basili et al., 1994):

Purpose: Study and analyze
Issue: The design and execution of A/B testing
Object: In software systems
Viewpoint: From the view point of researchers.

Concretely, we aim to investigate the subject of A/B testing, how A/B tests are designed and executed, and what the role is of stakeholders in the different phases of A/B testing. Finally, we also aim at obtaining insights in the research problems reported in the literature.

3. Methodology

This study uses the methodology of a systematic literature review as described in Keele et al. (2007). This methodology describes a rigorous process to review the literature for a topic of interest. The process ensures that the review identifies, evaluates, and interprets all relevant research papers in a reproducible manner. The literature review consists of three main phases: planning, execution, and synthesis. During planning a protocol is defined for the study (Quin et al., 2023), which

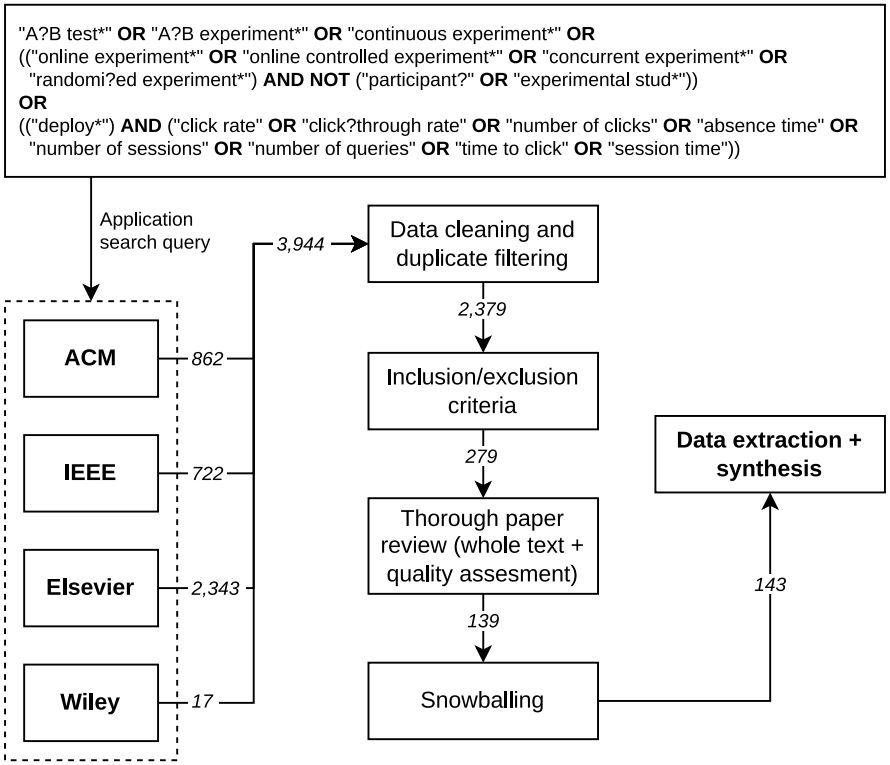

Fig. 2. Primary studies selected for the systematic literature review.

includes the motivation for the study, the research questions to be answered, sources to search for papers, the search string, inclusion- and exclusion criteria, data items to be extracted from the primary studies,[1] and analysis methods to be used. During execution the search string is applied as specified in the protocol, the inclusion and exclusion criteria are applied to identify the primary studies, and all the data items are extracted from these papers. Lastly, during synthesis the extracted data is analyzed and interpret to answer the research questions, and to obtain useful insights from the study.

We conducted the systematic literature review with four researchers. Further details on the process of the literature review (e.g. the roles the researchers play in the literature review) are summarized in the following sections. A complete description with the protocol, with all collected data and the data analysis are available at the study website (Quin et al., 2023).

3.1. Research questions

To realize the aim of this study ("Study and analyze the design and execution of A/B testing in software systems from the view point of researchers".), we put forward four research questions:

RQ1 What is the subject of A/B testing?
RQ2 How are A/B tests designed? What is the role of stakeholders in this process?
RQ3 How are A/B tests executed and evaluated in the system? What is the role of stakeholders in this process?
RQ4 What are the reported open research problems in the field of A/B testing?

With RQ1, we investigate the subject of A/B testing, i.e., the (part of the) system to which an A/B test is applied. Examples include A/B tests on program variables, application features, software components, subsystems, the system itself, and infrastructure used by the system. We also investigate the domains in which A/B testing is used.

With RQ2, we investigate what is defined and specified in A/B tests before they are executed in the system. We look at the metrics used, whether statistical methods are used in the experiments and if so which methods, and the tools used to conduct the experiments We also investigate which stakeholders are involved in this process and what is their role (e.g., users of the system influencing the tests that should be deployed, or architects deciding on which population the A/B tests should be run).

With RQ3, we investigate how A/B tests are executed in the system and the results are evaluated. More specifically, we look at the way in which data is collected for evaluation in the test, the evaluation of the A/B test itself (using the collected data and, if applicable, the result of a statistical test), and the use of the test results (e.g., decision about selection of target, input for maintenance, trigger for next test in a pipeline). We also explore the role stakeholders have during this process of A/B testing (e.g., operators deciding when to finish an experiment).

With RQ4, we identify open research problems in the field of A/B testing. The problems can be derived from descriptions of limitations of proposed approaches in the reviewed papers, open challenges, or outlines of future work on A/B testing.

3.2. Search query

We first identified a list of relevant terms for A/B testing from a number of known publications (King et al., 2017; Gui et al., 2015; Kohavi et al., 2009; Kohavi and Longbotham, 2017; Kharitonov et al., 2017; Drutsa et al., 2015b). We then identified and applied a gold standard (Zhang and Ali Babar, 2010) to tune the terms. For a detailed description of the relevant terms and application of the gold standard, we refer to the research protocol (Quin et al., 2023). Fig. 2 (top) displays the final search query after applying the gold standard.

[1] We use the term "research paper" to refer to papers that we considered for the application of inclusion and exclusion criteria in the SLR, and the term "primary study" for the research papers that we selected for data extraction.

3.3. Search strategy

The search query was executed October 2022. The search query was applied to the title and abstract of each paper in the sources (not case-sensitive). The automatic search was applied on four main digital libraries that publish research results on computer science in general and software engineering in particular,[2] resulting in 3944 papers, as shown in Fig. 2. After filtering duplicate papers and selecting only journal versions of extensions of conference versions, 2379 research papers are left for further processing.

3.4. Search process

After collecting the papers, we applied the following inclusion criteria:

- **IC1** Papers that either (1) have a primary focus on A/B testing (or any of its known synonyms) or (2) describe and apply (new) design(s) of A/B tests; for example introducing a proof-of-concept;
- **IC2** Papers that include an assessment of the presented A/B tests, either by providing an evaluation through simulation with artificial data or field data, or through running one or more field experiments in a real system;
- **IC3** Papers written in English.

We defined IC1 such that we only include works that are relevant to the posed research questions, i.e., it is essential that the work focuses on A/B testing or their design and evaluation. Note that IC1 includes papers that address and present solutions to known challenges in A/B testing. IC2 ensured that only papers are included that contain data related to the design and/or running of A/B tests. Lastly, we only included papers that are written in English with IC3. We manually applied IC1 and IC2 by closely inspecting the title and abstract of each collected paper. In case of doubt, the full text of the paper was consulted to gauge the satisfaction of both criteria. We applied IC3 automatically by filtering for English papers only in the digital libraries.

Besides the inclusion criteria above, we also applied the following exclusion criteria:

- **EC1** Papers that report (systematic) literature reviews, surveys (using questionnaires), interviews, and roadmap papers;
- **EC2** Short papers ($\leq$ 4 pages),[3] demos, extended abstracts, keynote talks, and tutorials;
- **EC3** Papers with a quality score ≤ 4 (explained in Section 3.5).
- **EC4** Papers that provide no or only a very brief description of the A/B testing design process or execution process.

EC1, EC2, and EC3 excluded papers that do not directly contribute new technical advancements, preliminary works that have not been fully developed yet, or works that are not of sufficient quality. In this literature review we focus on mature, state-of-the-art research in the field of A/B testing to answer the research questions. EC4 excluded works that do not contain essential information to answer the research questions.

Papers that satisfied all inclusion criteria and none of the exclusion criteria were included as primary studies in the literature study. The application of inclusion and exclusion criteria to the titles and abstracts of the research papers resulted in 279 papers. A thorough reading of the papers further reduced the number of papers to 139. In addition to the research papers retrieved via the search string and filtered by applying inclusion/exclusion criteria, we applied snowballing on the cited works of these papers to capture potentially missed papers. With snowballing we discovered 4 additional papers, bringing the final number of primary studies to 143, as shown in Fig. 2.

[2] We have not included Springer as that library does not support automatic search of the title and abstract of studies.

[3] Papers published in the *Lecture Notes in Computer Science* format with < 8 pages are also considered short.

Table 1
Collected data items to answer the research questions.

Identifier	Data item	Purpose
D1	Authors	Documentation
D2	Year	Documentation
D3	Title	Documentation
D4	Venue	Documentation
D5	Publication year	Documentation
D6	Paper type	Documentation
D7	Authors sector	Documentation
D8	Quality score	Documentation
D9	Application domain	RQ1
D10	A/B target	RQ1
D11	A/B test type	RQ2
D12	Used metrics	RQ2
D13	Statistical methods employed	RQ2
D14	Role of stakeholders in the experiment design	RQ2
D15	Additional data collected	RQ3
D16	Evaluation method	RQ3
D17	Use of test results	RQ3
D18	Role of stakeholder in experiment execution	RQ3
D19	Open problems	RQ4

3.5. Data items

To be able to answer the research questions, we extract the data items listed in Table 1. For each data item we provide a detailed description.

D1-5: Authors, year, title, venue, and publication year used for documentation purposes.

D6: The type of paper. Options include: focus paper (focus on A/B testing itself, i.e., modifications, suggestions, or enhancements to the A/B testing process), or applied paper (application and evaluation of A/B testing in real software systems).

D7: The sector of the authors of the primary study used for documentation (based on the author's affiliation). Options include Fully academic, Fully industrial, and Mixed.[4]

D8: A quality score for the reporting of the research (Mahdavi-Hezavehi et al., 2017). The quality score is defined on the following items: *Problem definition of the study*, *Problem context (relation to other work)*, *Research design (study organization)*, *Contributions and study results*, *Derived insights*, *Limitations*. Each item is rated on a scale of three levels: explicit description (2 points), general description (1 point), or no description (0 points). Therefore, the quality score is defined on a scale of 0 to 12 (Madeyski et al., 2014).

D9: The application domain that is used in relation to A/B testing in the primary study. Initial options include E-commerce, Telecom, Automotive, Finance, Robotics. Further options were be derived during data collection.

D10: The target of A/B tests describes the element that is subject of A/B testing. Initial options include an algorithm, a user interface, and application configurations. Further options were derived during data collection.

D11: The type of A/B test corresponding to the number of A/B variants and the way in which they are tested. Initial options include Single (classic) A/B test, Single multivariate A/B test, Manual sequence of classic A/B tests, Manual sequence of multivariate

[4] Academic refers to affiliations that are eligible to graduate master and/or Ph.D. students.

A/B tests, Automated sequence of classic A/B tests, Automated sequence of multivariate A/B tests. Additional options were derived during data collection.

D12: The metrics that are used in the A/B tests. Initial options include Click rate, Click-through rate, Number of clicks, Number of sessions, Number of queries, Absence time, Time to click, Session time. Additional options were derived during data collection.

D13: The statistical method that is employed to evaluate the data obtained through the A/B test, if any. Initial options include Student test, Proportional test, No statistical test. Further options were derived during data collection.

D14: The role of stakeholders in the experiment design. Initial options include Determining A/B test goal/hypotheses, Determining A/B test duration, Tuning A/B test variants. Further options are derived during data collection.

D15: Additional data that is collected during the execution of an A/B test (in addition to direct or indirect A/B metric data). Examples include User geo-location, Browser type, Timestamps of invocations or requests. Further options are derived during data collection.

D16: The evaluation method used in the primary study.[5] Initial options include Illustrative example, Simulation, Empirical evaluation.

D17: The use of the test results gathered from A/B tests. Examples include Subsequent A/B test execution, Subsequent A/B test design, Feature rollout, Feature development. Further options are derived during data collection.

D18: The role of stakeholder in the process of executing A/B tests. Initial options include A/B test alteration (adjusting individual A/B tests), A/B test triggering (starting subsequent A/B tests manually), A/B test supervision (monitoring A/B tests execution), No involvement, Unspecified. Further options are derived during data collection.

D19: Reported open problems. Open problems are derived from the reported challenges, limitations, and threats to validity. Options are derived during data collection.

4. Results

We start with the demographic information about the primary studies. Then we zoom in on each of the research questions.

4.1. Demographic information

Demographic information is extracted from data items Publication year (**D5**), Paper type (**D6**), Authors sector (**D7**), and Quality score (**D8**).

Fig. 3 shows the trend of number of publications of the primary studies over the years. We see a steady increase in the number of publications on the topic of A/B testing starting from the year 2009. We also observe a peek in the number of publications in 2017, counting 26 publications in total.

Of the 143 primary studies, 92 (64.3%) have a focus on A/B testing itself, while 51 (35.7%) apply A/B testing or use it for evaluation purpose, see Table 2.

A majority of 74 primary studies (51.7%) have industrial authors, see Table 3. Forty-three studies (30.0%) have a mix of industry and academic authors, and 26 studies (18.3%) are from academic authors only.

Table 2
Paper types of the primary studies.

Type	Number of occurrences
Focus	92
Applied	51

Table 3
Author backgrounds of the primary studies.

Background	Number of occurrences
Academic	26
Industry	74
Mixed	43

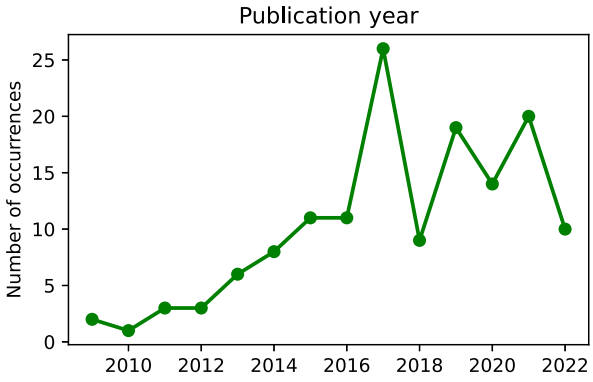

Fig. 3. Publication years of the primary studies.

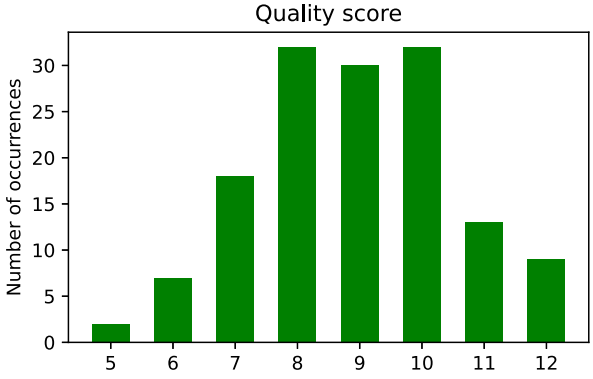

Fig. 4. Quality scores of the primary studies.

Fig. 4 shows the distribution of quality scores with an average of 8.92 [±1.60]. This shows that the reporting of the research in the primary studies is of good quality. Since all papers passed the threshold of 4, none of the papers had to be excluded for the extraction of data to answer the research questions.

4.2. RQ1: What is the subject of A/B testing?

To answer this research question, we look at the following data items: Application domain (**D9**), and A/B target (**D10**).

[5] We distinguish data retrieved from empirical evaluation in a live system from data retrieved from simulation or an illustrative example to provide targeted insights into the execution of A/B tests during data analysis of the SLR.

Table 4
Identified application domains for A/B testing.

Application domain	Number of occurrences
Web	38
Search engine	36
E-commerce	27
Interaction	23
Finances	16
Transportation	4
Other	8
N/A	9

Application domain. Table 4 lists the application domains of the primary studies. The average number of domains is 1.13 (133 primary studies applied A/B testing in one domain, three studies in two domains, six studies in three domains, and one study in four domains). Nine studies do not mention any domain. We observe that the most popular application domain is the Web (38 occurrences). Typical examples are social media platforms, such as Facebook (Liou and Taylor, 2020) or LinkedIn (Xu et al., 2014), news publishers (Yoneda et al., 2019; Fabijan et al., 2019), and multimedia services, such as movie streaming at Netflix (Amatriain, 2013). The second most popular domain is search engines (36 occurrences), with studies conducted at Yandex (Drutsa et al., 2017b,a), Bing (Dmitriev et al., 2016; Machmouchi et al., 2017), Yahoo (Akase et al., 2021; Tagami et al., 2014), among others. A/B testing is also actively applied in E-commerce (27 occurrences), with examples from retail giant Amazon (Ermis et al., 2020), the fashion industry (Chakraborty et al., 2021), and C2C (consumer-to-consumer) businesses, such as Etsy (Ju et al., 2019) and Facebook marketplace (Ha-Thuc et al., 2020). Next we observe the application of A/B testing in what we group under "interaction" (23 occurrences), with digital communication software, such as Snap (Xie et al., 2021) and Skype (Fabijan et al., 2019), user-operating system interaction (Gupchup et al., 2018; Fabijan et al., 2018), and application software, such as an App store (Dai et al., 2021) and mobile games (Yang et al., 2018). Lastly, we note the financial application domain (16 occurrences), including studies at Yahoo finance (Zhao et al., 2017) and Alipay (Cai et al., 2021), transportation (4 occurrences) at for instance Didi Chuxing (Fu et al., 2020). Other domains are education (3 occurrences) (Reza et al., 2021) and robotics (2 occurrences) (Mattos et al., 2017b), among others.

A/B target. The target of the A/B test denotes the element that is subject to testing and of which (at least) two variants are compared. Table 5 lists the A/B targets we identified from the primary studies, with a description and examples for each. The average number of A/B targets is 1.26 (89 primary studies applied A/B testing to one element, 13 studies to two elements, and 8 studies to three elements). Note that studies with more than one A/B target typically apply these in multiple experiments. The dominating targets of A/B testing are algorithm, visual elements, and workflow/process that together make up 85.6% of all A/B targets reported in the primary studies. Notable, 33 primary studies did not specify a particular A/B target, for example using datasets from two prior A/B tests in the paper's evaluation without clarifying the details of these tests (Xie et al., 2018).

Application domain vs A/B target. We can now map the application domains with the targets of A/B testing. This analysis provides insights into which elements or components are typically the subject of A/B testing in particular domains, or alternatively which A/B targets remain unexplored in particular domains. Table 6 presents this mapping. We highlight a number of key observations:

- A/B testing of algorithms is applied across all application domains and for all major domains it is the primary target of A/B testing. Commonly tested algorithms include feed ranking algorithms for social media websites, recommendation algorithms for news/multimedia websites, search ranking algorithms for search engines, and advertisement serving algorithms both in the Web and search engine application domains.
- A/B testing of visual elements is particularly popular for search engines (16 studies) compared to other application domains such as Web (with only 6 studies). Typical examples include changes to font color of search engine results (Kohavi et al., 2014) and changing the position of advertisements on the result page (Metrikov et al., 2014).
- Workflow and process elements as A/B target are commonly applied across the major domains. This target is particularly popular for the Web and E-commerce (with 8 and 7 studies, respectively). Typical examples are changes to the process in which best-performing advertisements are determined in JD's advertisement platform, China's largest online retailer (Wang et al., 2022), and changes to the order assignment policy for on-demand meal delivery platforms (Li et al., 2021c).
- For the Web and search engines, all types of A/B targets are applied. The main focus for the Web is on algorithms and workflow/processes, while the focus for search engines is on algorithms, visual elements, and back-end. For the Web, we notice only a single primary study with back-end as A/B target. This study targets different microservice configurations in A/B testing in order tune individual microservices for performance improvements (Sriraman et al., 2019). On the other hand, for search engines, we only noted three primary studies that target a workflow or process in A/B testing. One study evaluated a change of wording in digital advertisements (Blask, 2013), one study evaluated a change in advertisement strategies (Gupta and Chokshi, 2020), the last study evaluated the option to pay for "sponsored search" (to prioritize search results) (Blask et al., 2011).
- For e-commerce, we noticed that A/B testing is mainly used to test changes to ranking and recommendation algorithms, and to processes such as virtual assistants. Notably, we only identified a single primary study that evaluated changes to the user interface (Li et al., 2022).
- A/B testing for back-end optimizations was identified to be most common for search engines, while we did not identify a paper in e-commerce and finances domain where A/B testing was used for back-end changes.

Research question 1: What is the subject of A/B testing? The main targets of A/B testing are algorithms, visual elements, workflow and processes, and back-end features. A/B testing is commonly applied in the domains of Web, search engines, e-commerce, interaction software, and finances. Algorithms are consistently tested across these domains. Visual elements are predominantly evaluated in search engines, and counter-intuitively not in e-commerce. Workflow and processes are popular A/B targets in the Web and e-commerce domains. On the other hand, back-end features such as server performance are popular targets for search engines.

4.3. RQ2: How are A/B tests designed? What is the role of stakeholders in this process?

To answer the second research question, we look at the following data items: A/B test type (**D11**), Used metrics (**D12**), Statistical methods employed (**D13**), and Role of stakeholders in the experiment design (**D14**).

4.3.1. Design of A/B tests

To answer the first part of RQ2 (How are A/B tests designed?), we take a deeper look at the design of the A/B tests, focusing on the type of A/B tests, A/B metrics, and statistical methods used in the A/B tests.

Table 5
Identified A/B targets, with description.

A/B target	Description	Number of occurrences
Algorithm	Updated version of an algorithm such as a recommendation algorithm (Yoneda et al., 2019), a search ranking algorithm (Kharitonov et al., 2015), or an ad serving algorithm (Barajas et al., 2012b).	58
Visual elements	Change to visual components such as updates to a website layout (Brown et al., 2016) or a general user interface update (Dimmery et al., 2019).	33
Workflow/process	Alteration to the workflow of an application, e.g. the addition of a feedback button to a dashboard (Liu et al., 2019), or a change in a user workflow, e.g. the process of a virtual assistant tool (Koukouvis et al., 2016).	28
Back-end	Optimization of a software component that is not directly visible to the user, such as testing server optimizations (Poyarkov et al., 2016) or adjusting application parameters for better performance (Fabijan et al., 2019).	11
New application functionality	Newly introduced functionality, such as a new widget on a web-page (Chen et al., 2019) or additional content that is presented to the user after performing a search query (Machmouchi et al., 2017).	6
Other	This category comprises three other A/B targets: different timing and content of emails sent (Ye et al., 2022), varying educational resources presented to the user (Reza et al., 2021), and the page configuration of a web-site (Vasthimal et al., 2019).	3
Unspecified	The target of the A/B test was not specified in the study.	33

Table 6
Application domain ×A/B target.

A/B target / *Application domain*	Algorithm	Visual elements	Workflow /process	Back-end	New app. func.	Other
Web	17	6	8	1	3	0
Search engine	17	16	3	7	2	0
E-commerce	10	2	7	0	0	1
Interaction	5	6	2	2	1	0
Finances	7	2	4	0	1	0
Transportation	2	0	0	1	1	0
Other	2	1	3	0	0	2

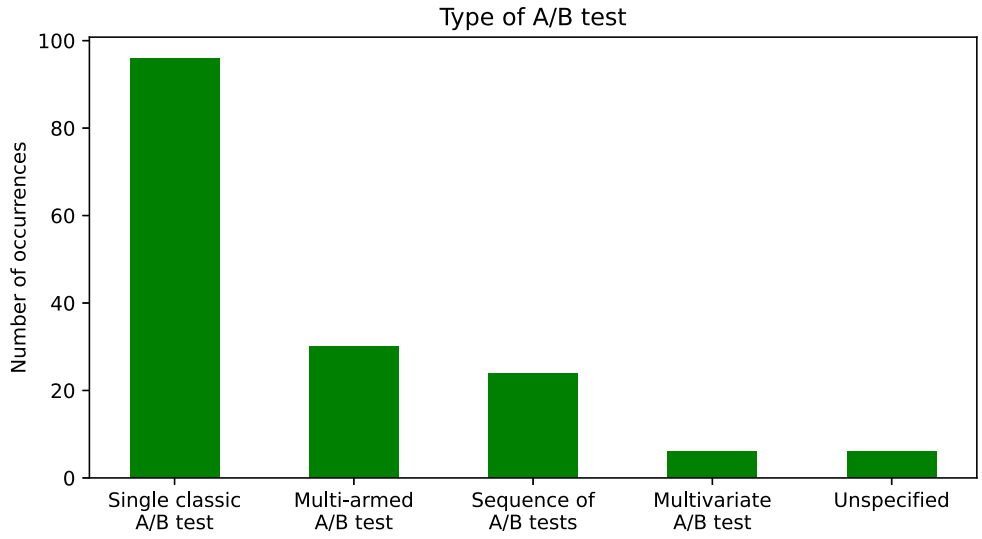

Fig. 5. Identified A/B test types.

A/B test type. The type of A/B tests include single classic A/B tests with two variants, A/B test composed of more than two variants (denoted as multi-armed A/B tests), multivariate A/B test where combinations of elements are tested in one A/B test, and sequences of all the these types. Fig. 5 shows the frequencies of these different A/B test types extracted from the primary studies.

Overall, we identified 156 occurrences of A/B test types, i.e., an average of 1.13 occurrences per primary study (124 studies considered a single type of A/B test, 17 studies considered two types, and one study considered three test types). The majority of the primary studies employed single classic A/B testing with a control variant and a treatment variant (96 occurrences). These standard test is used to test a variety of targets. The second most common type of A/B test is a multi-armed A/B test (30 occurrences). This type of test is composed of more than two variants under test; for example one control variant as baseline and three treatment variants with a distinct version each. These tests are commonly used to evaluate multiple versions of a recommendation algorithm, e.g., Shi et al. (2015), Szpektor et al. (2013), and to test

Table 7
Identified A/B metrics.

A/B metric	Number of occurrences
Engagement metrics	226
Click metrics	83
Monetary metrics	64
Performance metrics	50
Negative metrics	34
View metrics	21
Feedback metrics	17

Table 8
Statistical methods employed during A/B testing.

Statistical methods employed	Number of occurrences
Hypothesis - equality	57
Hypothesis - equality (concrete method unspecified)	39
Bootstrapping	11
Hypothesis - inference	8
Goodness of fit	8
Estimator	8
Correction method	7
Hypothesis - independence	5
Regression method	2

different advertisement serving algorithms, e.g., Ueoka et al. (2020). The third most common type of A/B test is a sequence of classic A/B tests (24 occurrences). Examples here include the comparison of multiple variants in a manually executed sequential style (as opposed to a multi-armed A/B test where all variants are deployed simultaneously) (Fairstein et al., 2022), manually testing multiple iterations of machine learning algorithms sequentially (Li et al., 2021a), and automatically executing a sequence of A/B tests to handle controlled feature release in DevOps (Schermann et al., 2016). The last identified A/B test type is multivariate A/B test (6 occurrences). This type of test evaluates various combinations of multiple features. As opposed to a multi-armed A/B test, a multivariate A/B test enables testing variants of more than a single feature in a singular A/B test. An example is the comparison of different combinations of varying GUI elements (Dimmery et al., 2019).

A/B metrics. Table 7 lists the A/B metrics that we extracted from the primary studies. In total, 495 occurrences of A/B metrics were reported in the primary studies. With a total of 200 experiments spread over 143 studies, this gives an average of 2.11 metrics per experiment[6] (ranging from 1 to 8 metrics per experiment). The most common group of A/B metrics are engagement metrics (226 occurrences) that refer to the number of conversions,[7] number of user sessions, time users are present on the website, and metrics related to the usage of the application or website (e.g. number of posts rated, number of bookings made).[8] The second largest group are click metrics (83 occurrences). Examples include number of clicks, clicks per query, and good click rate.[9] The third group of A/B metrics we identified are metrics related to monetization, i.e., revenue and cost (64 occurrences). Examples include number of purchases, order value, revenue per e-mail opening, and advertisement cost. The next group are performance metrics (50 occurrences). Examples include a simple response time of an application, bandwidth used, end-to-end latency, or playback delay of audio. The remaining groups are metrics that track unwanted effects in the A/B tests (34 occurrences, e.g. abandonment rate or number of un-subscriptions), views (21 occurrences, e.g. number of page views or number of product views), and user feedback (17 occurrences, e.g. number of customer complaints or verbatim feedback).

Statistical methods. Table 8 groups the types of statistical methods used for A/B tests in the primary studies. The most commonly used statistical method are hypothesis tests that test for equality (96 occurrences in total). The main test used in this group is a student t-test, e.g. Gupta and Chokshi (2020), Goswami et al. (2015). Other tests in this group are the Kolmogorov–Smirnov test, e.g., Shariat et al. (2017), Mann–Whitney test, e.g., Satyal et al. (2019), and Wilcoxon signed-rank test, e.g., Vanderdonckt et al. (2019). Out of the 94 occurrences of this type of hypothesis test, 37 primary studies did not report the concrete test used in the analysis of the results.[10] The second most commonly used method is bootstrapping (11 occurrences). This method constructs multiple datasets by resampling the original dataset (Drutsa et al., 2017b). The newly constructed datasets are then typically used for equality hypothesis testing. The key benefit of this technique is the sensitivity improvements gained in the analysis of the results. However, a big drawback of the technique is that it is computationally expensive, especially for larger datasets (Liu et al., 2019). The third mostly commonly used statistical method is a hypothesis test that tests for inference, goodness of fit, and custom estimators (all three with 8 occurrences). Examples of inference hypothesis tests include using Bayesian analysis approach to ensure multiple simultaneously running experiments do not interfere (Kohavi et al., 2013), and a Bayesian approach to infer the causal effect of running ad campaigns (Barajas et al., 2012a). Examples of goodness of fit methods include sequential testing methods that are based on likelihood ratio tests (Ju et al., 2019), and a Wald test (Jiang et al., 2020). An example of a custom estimator includes an estimator that takes variance into account to determine the confidence interval of the A/B test (Liou and Taylor, 2020). The remaining groups are correction methods (7 occurrences) with e.g. Bonferroni correction (Yu et al., 2020); hypothesis tests for independence (5 occurrences), containing χ^2 tests (Tagami et al., 2014); and regression methods (2 occurrences), e.g. CUPED (Duan et al., 2021).

4.3.2. Role of stakeholders

To address the second part of RQ2 (What is the role of stakeholders in the design of A/B tests?), we analyze the role stakeholders play in the design of A/B tests.

Roles of stakeholders. Table 9 lists the different roles of stakeholders in the design of A/B tests that we extracted from the primary studies, associated with tasks, descriptions and examples. We identified three main roles: concept designer (128 occurrences), experiment architect (113 occurrences), and setup technician (31 occurrences). The role *Concept designer* consists of conceptualizing new ideas for A/B testing. The role of *Experiment architect* consists of calibrating technical parameters of the experiment such as the experiment duration. The role of *Setup technician* consists of taking the necessary steps required to allow the execution of the A/B test. The top task of the concept designer is designing and tuning variants of A/B tests (67 occurrences). The top task of the experiment architect is determining the duration of A/B tests (61 occurrences). Finally, the main task of the setup technician is performing post-design activities of A/B tests (25 occurrences).

[6] We excluded experiments and corresponding metrics of primary studies that analyzed a large number of previously conducted A/B tests.

[7] A conversion is a desired action taken in the A/B test.

[8] Note that some of the primary studies do not specify explicitly the A/B metrics due to business sensitivity. Based on the available information in the study, we have included these in general engagement metrics.

[9] Good clicks are described as clicks that are meaningful during the search query session (Borisyuk et al., 2021).

[10] However, these studies did report p-values alongside the results, or explicitly refer to confidence intervals and statistically significant results of the A/B tests.

Table 9
Roles and tasks of stakeholders in the design of A/B tests (Occ short for number of occurrences).

Role	Task	Task description	Occ.
Concept designer (128)	Design and tune variants	Designing and tuning the variants to test. Examples are tweaking the A/B variants (Shi et al., 2015), or designing A/B variants for different kind of populations (e.g., old vs new users) (Borodovsky and Rosset, 2011).	67
	Determine goal or hypothesis	Formulating the goal or hypothesis of the A/B test itself. Examples include the specification of a goal to find the better performing news selection algorithm (Eckroth and Schoen, 2019) or the specification of a pre-determined hypothesis for the A/B test (Aiello et al., 2016).	48
	Perform pre-design actions	Actions that are taken before designing the A/B test. Examples include providing motivation for A/B tests (Vasthimal et al., 2019) or performing offline A/B tests before moving to online A/B testing (Gruson et al., 2019).	13
Experiment architect (113)	Determine duration	Determining the duration of the A/B test. Examples include choosing a fixed experiment duration (e.g., 1 week) (Aiello et al., 2016) or via an explicit expiration date (Li et al., 2019).	61
	Determine population assignment	Determining the population that should take part in the A/B test. Examples include a simple 50/50 split of all users (Wu and Grbovic, 2020), an assignment where the target population is determined over a two week period (Yang et al., 2018), or an assignment where network effects have to be taken into account (Li et al., 2021c).	52
Setup technician (31)	Perform post-design actions	Actions that are taken after completing the design of the A/B test. Examples include performing A/A testing prior to running the A/B test (Zhao et al., 2017; Dai et al., 2021), validation of the A/B test design (Liu et al., 2019), or scheduling the execution of the A/B test (Vasthimal et al., 2019).	25
	Perform metric analysis and initialization	Analyzing and potentially initializing metrics for the A/B test. An example consists of instantiating a custom A/B utility metric with negative and positive weights tied to user's actions during a search session (Machmouchi et al., 2017).	6

Table 10
Tasks of stakeholders ×A/B test type.

Test type (total occ.) / *Task*	Single classic A/B test (95)	Multi-armed A/B test (30)	Sequence of A/B tests (24)	Multivariate A/B test (6)
Design and tune variants	33	22	13	2
Duration	46	9	11	2
Population assignment	38	7	8	2
Goal/hypothesis	27	17	8	2
Post-design actions	12	1	5	0
Pre-design actions	7	4	2	1
Metric analysis/init.	5	1	0	0

4.3.3. Cross analysis A/B test design

We discuss two mappings of data items: The role stakeholders take in the design of A/B tests versus A/B test type; and the A/B metrics used in experiments versus the statistical methods employed.

Tasks of stakeholders vs A/B test type. The mapping of stakeholder's tasks in the design of A/B tests across types of A/B tests is shown in Table 10. We observe the following:

- The primary tasks of stakeholders across all types of A/B tests are the design and tune of variants, determining the duration of experiments, the population, and the goal or hypothesis. These numbers confirm that these are essential design tasks for any A/B test.
- A majority of the studies that use multi-armed A/B testing and sequence of A/B tests report the design and tuning of variants as important stakeholder task (22 and 13 occurrences respectively). Since these types of tests involve multiple variants under test, the studies often specify more details about the variants and the reasoning behind choosing which variants to test.
- Determining the goal or hypothesis for A/B testing is frequently mentioned for multi-armed A/B tests (17 occurrences). In contrast to conventional two-variant A/B testing that typically involves a control variant and an altered variant aimed at improving the control variant, multi-armed A/B tests involve more than two variants, so practitioners often formulate hypotheses regarding the potential performance of each variant.
- Post-design actions are more often reported for sequences of A/B tests (5 instances). For instance, one primary study mentions modeling the sequence of A/B tests (Schermann et al., 2016), another study mentions determining the success condition of

Table 11
Statistical methods ×A/B metrics (H short for hypothesis)

Method \ *Metric*	Engag.	Click	Monetary	Negative	Perf.	View	Feedback
H - equality	31	14	7	10	4	7	2
H - equality (unsp.)	25	13	8	8	11	5	5
Bootstrapping	9	2	2	3	3	1	1
H - inference	5	1	0	0	1	0	0
Goodness of fit	5	1	2	1	0	0	0
Correction method	4	1	1	1	2	0	1
Estimator	5	2	2	1	0	1	0
H - independence	2	2	3	0	0	1	0
Regression method	1	1	1	0	0	1	0

the A/B tests before executing them (Tamburrelli and Margara, 2014), and another study refers to providing an outcome range of the A/B tests (Tang et al., 2010).
- Only a few primary studies report pre-design actions and metrics analysis and initialization, independently of the type of A/B test.

A/B metrics vs statistical methods used. The statistical methods used across different types of A/B metrics are shown in Table 11.

- Engagement metrics and click metrics are used across all types of statistical methods.
- The concrete method used for hypothesis testing of equality is often not specified across all types of A/B metrics. For monetary and performance metrics in particular, a majority of studies do not mention the concrete hypothesis testing method (8 and 11 occurrences, respectively). This might be due to the sensitivity in reporting results for these types of metrics.
- Negative metrics are primarily used for hypothesis equality tests (10 and 8 occurrences for hypothesis equality and hypothesis equality no method specified respectively).
- Hypothesis method for independence is most frequently used for the monetary metrics, yet, the use is uncommon (3 instances).
- The use of feedback metrics is also uncommon and if used, the specific statistical method used is not reported (5 occurrences).

Research question 2: How are A/B tests designed? What is the role of stakeholders in this process? The primary type of A/B test is a single classic A/B test, followed by multi-armed A/B tests and sequence of A/B test. Engagement metrics are the dominating type of A/B metrics used in A/B testing. Other prominent A/B metrics include click, monetary, and performance metrics. Hypothesis testing for equality is by far the most commonly used statistical method used in A/B testing. Remarkable, about 40% of these studies that test on equality do not specify the concrete method they use for that. Stakeholders have two main roles in the design of A/B tests: concept designer and experiment architecture. Less frequently reported is a third role of setup technician.

4.4. RQ3: How are A/B tests executed? What is the role of stakeholders in this process?

To answer research question 3, we look at the following data items: Data collected (**D15**), Evaluation method (**D16**), Use of test results (**D17**), and the Role of stakeholders in the experiment execution (**D18**).

4.4.1. Execution of A/B tests

To address the first part of RQ3 (How are A/B tests executed?), we analyze the data collected during A/B tests, the evaluation methods used, and the use of A/B tests.

Table 12
Data collected for the A/B tests.

Data collected	Number of occurrences
Product/system data	49
User-centric data	26
Spatial–temporal data	20
Secondary data	6

Table 13
Evaluation method used in the primary studies.

Evaluation method	Number of occurrences
Empirical evaluation	101
Simulation based on real empirical data	27
Simulation	15
Illustrative example	10
Case study	5
Theoretical	1

Data collected. Table 12 lists the classes of data collected during the execution of A/B tests. We identified four types of data. Product or system data is most commonly reported in the primary studies (49 occurrences). This data class includes the type of browser used by the end-user, the operating system of the end-user, hardware-specific information of the device used to interact with the application, and general information related to usage of the system (e.g. tracking information about item categories of products in an e-commerce application, and types of search queries processed during the A/B test). Second most popular is user-centric data (26 occurrences). This class contains data related to how the end-user interacts with the system as well as personal information of end-users. Examples include scrolling characteristics of users on a web application, the navigation history of end-users, user feedback, and using age or current occupation of the end-user during analysis. The third most commonly reported class is spatial–temporal data (20 occurrences) that groups data related to geographic location and time-related data. Examples include timestamps of requests to an application, the creation date of accounts that take part in the A/B test, and spatial information such as the country and region of end-users. Lastly, a few primary studies report the use of secondary data (6 occurrences). Data in this class correspond to A/B metrics that do not serve as main evaluation metrics for A/B tests. Examples are the number of clicks or page views that are used for additional analysis after conducting the A/B tests.

Evaluation method. Table 13 summarizes the identified evaluation methods. The vast majority of primary studies provide results from an empirical evaluation (101 occurrences), i.e., executing A/B tests in live systems. A substantial number of studies use historical data from previously conducted A/B tests to simulate new A/B tests (27 occurrences), while a handful of studies (15 occurrences) use simulations without historical data as their evaluation method. Lastly, a few studies use illustrative examples (10 occurrences), case studies (5 occurrences),

Table 14
Use of test results gathered from A/B test execution.

Use of test results	Description	Occur.
Feature selection	The results of the A/B test are used to determine which variant presents an improvement to the application. Examples include selecting a new version of a ranking algorithm (Pelleg et al., 2016; Chen et al., 2019) or a recommendation algorithm (Freno, 2017), and selecting a different visual design (Almeida and Casais, 2022).	71
Feature rollout	The results of the A/B test are used to determine if the rollout of a feature should be continued or halted, as for example outlined by practitioners at Microsoft (Xia et al., 2019; Duivesteijn et al., 2017).	24
Continue feature development	The results of the A/B test are used as a driving force for further feature development, e.g. fine-tuning newly proposed A/B metrics based on periodicity patterns after obtaining promising results (Drutsa et al., 2017a), and further developing personalization methods (Akase et al., 2021).	17
Subsequent A/B test design	The results of the A/B test are used for future A/B test design, for example suggesting alternative A/B variants to test in future A/B tests (Koukouvis et al., 2016), and designing a new A/B test to further test the quality of an A/B metric prediction model[a] (Peska and Vojtas, 2020).	15
Validation effectiveness of A/B testing process	The results of the A/B test are used to demonstrate the effectiveness of the newly proposed or improved A/B testing approach by the authors. Examples include evaluating a newly proposed counterfactual framework to run seller-side A/B tests in two-sided marketplaces (Ha-Thuc et al., 2020), and the validation of a new statistical methodology for continuous monitoring of A/B tests (Johari et al., 2017).	12
Validation of a research question	A/B testing is used to validate a research question put forward by the authors. One example consists of investigating the hypothesis under which circumstances companies should pay for advertising in search engines (Blask et al., 2011).	10
Bug detection/fixing	The results of the A/B test are used to detect potential bugs or validate bug fixes, e.g. probing for data quality issues in A/B tests of ML models to uncover potential bugs (Li et al., 2021a).	5
Subsequent A/B test execution	The results of the A/B test are used to execute subsequent A/B tests, e.g. using the results of A/B tests to automatically determine which subsequent A/B tests to execute (Tamburrelli and Margara, 2014).	4
Unspecified	The use of the test results was not specified in the study.	26

[a] The prediction model was trained on data from the previously conducted A/B tests.

and a single primary study provides a theoretical evaluation (Metrikov et al., 2014).

Use of test results. Table 14 lists the use of test results extracted from the primary studies. Use of test results refers to what stakeholders do with the obtained data and analyses of A/B tests, such using the results to design additional A/B tests. As the table shows, the main usages of A/B test results are the selection and rollout of a feature (71 and 24 occurrences respectively). A number of studies aim at validating the effectiveness of the A/B testing process itself (12 occurrences). The use of test results to trigger a subsequent A/B test seems not very well explored (4 occurrences).

4.4.2. Role of stakeholders

To address the second part of RQ3 (What is the role of stakeholders in this process?), we analyze the role of stakeholders in A/B test execution.

Roles of stakeholders. Table 15 lists the different role of stakeholders in the A/B test execution we have extracted from the primary studies with associated tasks, a description and examples. We identified two main roles: experiment contributor (40 occurrences) and experiment assessor (37 occurrences). The role *Experiment contributor* consists of managing the A/B test execution. The role *Experiment assessor* consists of evaluating the A/B test results and potentially undertaking additional actions. The top task of the experiment contributor is experiment supervision (19 occurrences). The top task of the experiment assessor is experiment post-analysis (17 occurrences).

4.4.3. Cross analysis A/B test execution

We take a deeper look at two mappings of data items related to the execution of A/B tests: Use of test results with the tasks of stakeholders in the execution of A/B tests; and the evaluation method with the tasks of stakeholders in the execution of A/B tests.

Use of test results vs tasks of stakeholders in the execution of A/B tests. The first mapping we analyze relates to the use of test results and the tasks stakeholders undertake in the execution of A/B tests. The results are shown in Table 16. We highlight some key observations:

- Experiment supervision is applied regardless of the usages of test results. For feature rollout as a use of A/B test results, the task of experiment supervision is often mentioned. Supervision takes on a key task in this context to ensure that the rollout happens in a hazard-free manner (i.e., no harm is caused to users) (Xia et al., 2019; Chen et al., 2019).
- The task of experiment post-analysis is typically only reported for experiments that are fully complete (i.e., do not go through additional rounds of iteration). In the primary studies where the results of the A/B tests are used for subsequent A/B test design, no instances were identified where stakeholders take the task of performing post-analysis on the results of the experiments.
- For subsequent A/B test design, the task of experiment triggering is often mentioned. This is to be expected since the newly designed A/B tests also need to be executed. Additionally, A/B test termination is also mentioned often (e.g., terminating an experiment due to bad results Gupta et al., 2018).
- In the case of bug fixing and detection, stakeholders typically supervise experiments (either to detect possible bugs in the code or ensure the bugfix is effective) (Fabijan et al., 2017a), and trigger the experiments (i.e. launch an experiment explicitly to fix a known bug in the application) (Li et al., 2021a).

Evaluation method vs tasks of stakeholders in the execution of A/B tests. In addition, we analyze the tasks stakeholders undertake during the execution of A/B tests across the evaluation methods. This mapping is shown in Table 17. We highlight a number of key takeaways:

- All tasks that stakeholders undertake in the execution of A/B tests are widely encountered in the case of empirical evaluation.

Table 15
Identified roles and concrete tasks of stakeholders during in the execution of A/B tests.

Role	Task	Task description	Occ.
Experiment contributor (40)	Experiment supervision	Monitoring and closely following up on the execution of A/B tests (Tamburrelli and Margara, 2014; Dobaj et al., 2022).	19
	Experiment alteration	Altering aspects of the A/B test during its execution. Examples include adjusting the population assignment of the experiment (Dai et al., 2021), or adjusting the A/B variants themselves (Tang et al., 2010).	12
	Experiment termination	Stopping A/B tests when deemed necessary. Examples include manually stopping A/B tests when sufficient data is collected (Koukouvis et al., 2016), or stopping the experiment early when harm is observed (Kohavi et al., 2013).	9
Experiment assessor (37)	Experiment post-analysis	Various steps that are taken after analyzing the results of the A/B test. Examples include double checking results from executed A/B tests (Goswami et al., 2015), performing a deeper analysis of suspicious results (Fabijan et al., 2019), or performing bias reduction techniques on the retrieved data from the A/B tests (Liu et al., 2019).	17
	Experiment triggering	Starting the execution of (subsequent) A/B tests (Xia et al., 2019; Brown et al., 2016).	13
	Other	This category encompasses a few niche tasks, such as documenting the findings and learning from conducting the A/B test (Siqueira and de Paula, 2018), rerunning A/B tests (Machmouchi et al., 2017), or incorporating user feedback in the analysis of the A/B tests (Li et al., 2019).	7

Table 16
Use of test results ×Tasks of stakeholders in the experiment execution (“cont. feature dev.” is short for “continue feature development”, “val. eff.” is short for “validation of effectiveness”, and “val. of a RQ” is short for “validation of a research question”).

Use \ *Task*	Supervision	Post-analysis	Triggering	Alteration	Termination
Feature selection	8	11	6	8	4
Feature rollout	10	4	6	6	4
Cont. feature dev.	7	3	5	2	3
A/B test design	6	0	5	2	3
Val. eff. A/B testing	1	2	1	1	1
Val. of a RQ	1	1	0	1	1
Bug detection/fixing	4	0	3	2	2
A/B test execution	1	0	1	0	0

Table 17
Evaluation method ×Tasks of stakeholders in the test execution (“emp. sim.” short for “simulation based on real empirical data”, “ill.” short for “illustrative”).

Method \ *Task*	Supervision	Post-analysis	Triggering	Alteration	Termination	Other
Empirical	14	13	10	10	6	6
Emp. sim.	2	4	1	0	1	0
Simulation	1	1	1	0	0	0
Ill. example	2	0	1	2	2	1
Case study	0	0	0	0	0	0
Theoretical	0	0	0	1	0	0

- For the method of simulation based on real empirical data, the task of post-analysis is reported more often than any other task. An example is looking for outliers in the analysis of the results of A/B tests, and using historical experiments to confirm its effectiveness (He and Chen, 2017).
- Primary studies that use simulation as an evaluation method rarely specify the tasks stakeholders undertake in the execution of A/B tests. We hypothesize that, since simulations allow for a more controlled way of conducting A/B tests, the tasks stakeholders undertake after the design of A/B tests are not pertinent.

Table 18
List of identified open problems.

Open problem category	Open problem sub-category	Number of occurrences
Evaluation-related	Extend the evaluation	21
	Provide thorough analysis of approach	16
	Other evaluation-related	36
Process-related	Add process guidelines	9
	Automate process	7
Quality-related	Enhance scalability	7
	Enhance applicability	6

- The only stakeholder task reported for theoretical evaluation is experiment alteration (primary study (Metrikov et al., 2014)).

Research Question 3: How are A/B tests executed in the system? What is the role of stakeholders in this process? The main types of data collected during the A/B test execution relate to the product/system, users, and spatial–temporal aspects. The dominating evaluation method used in A/B testing is empirical evaluation, but a relevant number of studies also use simulation. A/B test results are primarily used for feature selection, followed by feature rollout, and continue feature development. (Automatic) subsequent A/B test execution is only used marginally. The main reported roles of stakeholders in A/B test execution is experiment contributor (with experiment supervisor as main task) and experiment assessor (with experiment post-analysis as main task).

4.5. RQ4: What are the reported open research problems in the field of A/B testing?

To answer research question 4, we analyze the results of data item Open problems (**D19**).

Table 18 present a categorization of open problems we have identified in the primary studies. For each category we devised concrete sub-categories of open problems. We elaborate on each type of open problem with illustrative examples.

4.5.1. Evaluation-related open problems

First, we established three sub-categories of open problems that are related to the evaluation of the proposed approach: (1) extensions to the evaluation of the approach presented in the primary study, (2) a more thorough analysis of the approach presented in the primary study, and (3) Other evaluation-related open problems in the primary study.

Extend the evaluation. Drutsa et al. (2017a) explore periodicity patterns in user engagement metrics, and its influence on engagement metrics in A/B tests. Moreover, the authors put forward new A/B metrics that take such periodicity patterns into account, resulting in more sensitive A/B test analysis. The authors evaluated the proposed metrics on historical A/B test data from Yandex, though they state that further evaluation of the approach could be carried out in different domains such as social networks, email services, and video/image hosting services. From a slightly different point of view, Barajas et al. (2012a) developed a technique to determine the causal effects of marketing campaigns on users, putting the focus on the campaign itself rather than only focusing on the design of advertisement media. The authors put forward specific guidelines on randomizing and assigning users to advertising campaigns, and provide a technique to estimate the causal effect the campaigns have on the users under test. As a point of future work, the authors posit a different evaluation question concerning what would have happened if the technique would have been applied to the whole population.

Provide thorough analysis of approach. An example of this category is mentioned by Peska and Vojtas (2020). The authors put forward a way of evaluating recommendation algorithms in small e-commerce applications both offline and online via A/B testing. The approach compares results of offline evaluation of recommendation algorithms with the results of online A/B testing of the algorithms. Moreover, the authors then used these data to build a prediction model to determine the promising recommendation algorithms more effectively due to the knowledge obtained from online A/B testing. As future work, the authors list that further work is necessary to verify the causality of an effect observed in the analysis of offline and online A/B testing data. In another primary study written by Madlberger and Jizdny (2021), the authors perform an analysis on the impact of social media marketing on click-through rates and customer engagement. To accomplish this, they run multiple social media marketing campaigns using A/B testing, evaluating hypotheses related to the impact of visual and content aspects of advertisements on the click rates of end-users. As future research, the authors report that a more comprehensive investigation is necessary to ascertain why some hypotheses in the study have been rejected.

Other evaluation-related. An example of other evaluation-related open problems is laid out by Gruson et al. (2019). The authors propose a methodology based on counterfactual analysis to evaluate recommendation algorithms, leveraging both offline evaluation and online evaluation via A/B testing. The approach comprises A/B testing recommendations to a subset of the population, and using the results of these tests to de-bias offline evaluations of the recommendation algorithm based on historical data. In regards to open problems, the authors mention exploring additional metrics for the approach, as well as potential improvements that can be made to the estimators they use in the approach. Another example is specified by Ju et al. (2019), who present an alternative to standard A/B testing with a static hypothesis test by putting forward a sequential test. Classically in A/B testing, the hypothesis of the test is tested after a fixed time and conclusions are made based on the final result. The sequential test put forward by the authors does not have a predetermined number of observations, rather at multiple points during the experiment the test determines whether the hypothesis can be accepted, rejected, or if more observations are required. For future work, the authors wish to support A/B/n experiments in their approach, as well as extending the procedure for data that follows a non-binomial distribution. In a final example, Gui et al. (2015) study ways of dealing with interference of network effects in the results of A/B tests. One of the fundamental assumptions of A/B testing is that users are only affected by the A/B variant they are assigned to. However, network effects can undermine this assumption do to interaction between users in the population. The authors demonstrate the presence of network effects at LinkedIn, and propose an estimator for the average treatment effect that also takes potential network effects into account. As a line of future research, the authors want to investigate ways of enhancing the approach such that it can deal with more real life phenomena.

4.5.2. Process-related open problems

Second, we established two sub-categories of open problems that are process-related: (1) guidelines to the A/B testing process, and (2) automation of aspects of the A/B testing process.

Add process guidelines. In an effort to provide more nuanced A/B testing guidelines in the e-commerce domain, Goswami et al. (2015) discuss controlled experiments to make decisions in the context of e-commerce search. Considerations such as how to prioritize projects for A/B testing for smaller retailers and how to conduct A/B tests during holiday time are left as open questions. A different primary study covering the benefits of controlled experimentation at scale is presented by Fabijan et al. (2017a). In this study, the authors present multiple examples of conducted A/B tests, and the corresponding lessons learned from these experiments. One of the listed open problems in the study relates to providing “guidance on detection of patterns between leading and lagging metrics”.

Automate process. Mattos et al. (2017b) present a step towards automated continuous experimentation. The authors put forward an architectural framework that accommodates the automated execution of A/B tests and automated generation of A/B variants. To validate the framework, an A/B test was conducted with a robot. One of the open challenges laid out in the study comprises the ability to automatically generate hypotheses for A/B tests based on the collected data. Duivesteijn et al. (2017) present A&B testing, an approach that leverages exceptional model mining techniques to target A/B variants to subgroups in the population under test. As opposed to deploying the best-performing variant of the A/B test, the authors put forward running both variants (if ample resources are available) and targeting specific variants to individual users based on their inferred subgroups. One of the potential avenues for future research consists of the development of a framework that would enable automated personalization of websites supported by A/B testing.

4.5.3. Quality-related open problems

Lastly, we established two sub-categories of open problems that are quality-related: (1) enhancing scalability of the proposed approach, and (2) enhancing the applicability of the approach.

Enhance scalability. One example of this is presented by Zhao et al. (2017). In order to obtain a causal explanation behind the results of A/B tests, the authors propose segmenting the population, and consequently analyzing the results of the A/B test in individual segments. For future work, the authors mention developing a more scalable solution that integrates the approach into their existing experimentation platform. To address online experimentation specifically for cloud applications, Toslali et al. (2020) introduce Jackpot, a system for online experimentation in the cloud. Jackpot supports multivariate A/B testing and ensures proper management of interactions in the cloud application during the execution of A/B tests. As a venue for future work, the authors mention ways of dealing with the limited scalability of multivariate experimentation due to the number of potential experiments increasing exponentially with the number of elements to be tested.

Enhance applicability. One such study explores A/B testing in the automotive industry (Liu et al., 2021). The study addresses concerns relating to the limited sample sizes A/B tests obtain due to the limited nature of participants that can take part in A/B tests in the industry. To overcome this hurdle, the authors provide specific guidelines for performing A/B testing and determining the assignment of users to either the control or treatment variant in the test. However, one limitation pertains to requiring pre-experimental data to ensure a balanced population assignment between both A/B variants. In an effort to increase sensitivity in A/B testing, Liou and Taylor (2020) propose a new estimator for A/B testing that takes variance of individual users into account. To realize this, pre-experiment data of individual users is analyzed and variances are computed. In order to validate the approach a sample of 100 previously conducted A/B tests were collected and analyzed using the new approach. A big limitation noted by the authors is that "a stronger assumption about the homogeneity of the treatment effect" is required in order for the approach to remain unbiased.

Research Question 4: What are the reported open research problems in the field of A/B testing? The most commonly reported open problems directly related to the proposed approach, in particular improving the approach, extending the approach and providing a thorough analysis. Other less frequently reported open problems relate to the A/B testing process, in particular adding guidelines for the A/B testing process, and automating the process. Finally, a number of studies report open problems regarding quality properties, specifically enhancing scalability and applicability of the proposed approach.

5. Discussion

In this section, we discuss a number of additional insights we obtained. We start with the research topics studied by the primary studies. Next, we look at environments and tools used for A/B testing. Then we report a number of opportunities for future research. We conclude with a discussion of threats to validity of the study.

5.1. Research topics

During data extraction of the 143 primary studies, we noted the general subject matters of the primary studies and categorized the primary studies along 7 research topics. Table 19 summarizes these 7 topics. Note that studies share overlapping topics. We briefly explain now each category and provide a few examples from the primary studies.

5.1.1. Application of A/B testing

The main focus of the primary study is the use and application of A/B testing as evaluation tool for the main subject matter of the study (e.g. evaluation new recommendation algorithm, interface redesigns, etc.[11])

5.1.2. Improving the efficiency of A/B testing

This topic is about improving the process of A/B testing by exploring ways of improving sensitivity in A/B testing data (Drutsa et al., 2015a; Poyarkov et al., 2016; Wu et al., 2022; Kharitonov et al., 2017), investigating sequential testing techniques to stop A/B tests as soon as reasonable (Ju et al., 2019; Kharitonov et al., 2015; Abhishek and Mannor, 2017), proposing techniques to detect invalid A/B tests[12] Chen et al. (2019), and using extra data such as periodicity patterns in user behavior to improve A/B testing (Drutsa et al., 2017a).

5.1.3. Beyond standard A/B testing

This topic is about techniques that go beyond standard A/B testing, such as the use of new types of A/B metrics (Xie et al., 2018; Duan et al., 2021; Machmouchi et al., 2017), use of counterfactuals in the evaluation of A/B tests[13] (Ha-Thuc et al., 2020; Rosenfeld et al., 2017), investigating ways of automating parts of the A/B testing process (Schermann et al., 2016; Mattos et al., 2017b,a; Tamburrelli and Margara, 2014), improving or altering the A/A testing process (He et al., 2022; Chen et al., 2017b), and investigating ways of combining offline- and online A/B testing (Gruson et al., 2019; Peska and Vojtas, 2020).

5.1.4. Concrete A/B testing problems

This topic includes studies that A/B testing in specific domains and specific types of A/B testing. Examples include A/B testing specifically in the e-commerce domain (Koukouvis et al., 2016; Goswami et al., 2015), network A/B testing or A/B testing in marketplaces (Li et al., 2022; Gui et al., 2015; Cai et al., 2021), A/B testing in the CPS domain with digital twins (Dobaj et al., 2022), or A/B testing for mobile applications (Lettner et al., 2013; Xu and Chen, 2016).

5.1.5. Pitfalls and challenges of A/B testing

This topic is about pitfalls related to conducting A/B testing (Esteller-Cucala et al., 2019; Kohavi et al., 2012; Dmitriev et al., 2017, 2016), or (particular domain-related) challenges related to A/B testing (Xu et al., 2015; Mattos et al., 2020; Liu et al., 2019).

[11] See data item *A/B target* in for specific references.

[12] Invalid refers to badly designed experiments or misinterpretation of the results retrieved from the experiment.

[13] Counterfactual analysis provides answers to the cause and effect of the treatment group and their corresponding outcomes, compared to what would have happened if the treatment would not have been applied.

Table 19
Research topics of primary studies.

Topic	Number of occurrences	Primary studies
Application of A/B testing	51	Yoneda et al. (2019), Okoshi et al. (2019), Kornilova and Bernardi (2021), Li et al. (2021c), Barajas et al. (2012b,a), Metrikov et al. (2014), Xu et al. (2021), Lalmas et al. (2015), Sun et al. (2018), Dai et al. (2021), Gomez-Uribe and Hunt (2016), Fu et al. (2020), Ermis et al. (2020), Borisyuk et al. (2021), Ye et al. (2022), Li et al. (2017), Fairstein et al. (2022), Ueoka et al. (2020), Tagami et al. (2014), Freno (2017), Wu and Grbovic (2020), Xu et al. (2014), Chen et al. (2017a), Silberstein et al. (2020), Agarwal et al. (2014), Aiello et al. (2016), Sajeev et al. (2021), Szpektor et al. (2013), Alfaro-Flores et al. (2021), Kumar and Hasteer (2017), Sriraman et al. (2019), Shi et al. (2015), Yang et al. (2018), Blask et al. (2011), Chakraborty et al. (2021), Almeida and Casais (2022), Madlberger and Jizdny (2021), Akase et al. (2021), Miikulainen et al. (2020), Eckroth and Schoen (2019), Kumar and Hosanagar (2017), Satyal et al. (2017), Pelleg et al. (2016), Brown et al. (2016), Ozawa et al. (2016), Qu and Zhang (2016), Wang and Hardtke (2015), Funk (2009), Aharon et al. (2019a), Young (2014)
Improving efficiency of A/B testing	21	Abhishek and Mannor (2017), Chen et al. (2019), Poyarkov et al. (2016), Wu et al. (2022), Budylin et al. (2018), Kharitonov et al. (2017), Drutsa et al. (2015b), Deng et al. (2021), Drutsa et al. (2015a), Kharitonov et al. (2015), Dimmery et al. (2019), Drutsa et al. (2017a,b), Liou and Taylor (2020), Lee and Shen (2018), Ju et al. (2019), Blask (2013), He and Chen (2017), Deng et al. (2013), Deng (2015), Feit and Berman (2019)
Beyond standard A/B testing	18	Xie et al. (2018), Deng et al. (2014), Johari et al. (2017), Duan et al. (2021), Ha-Thuc et al. (2020), He et al. (2022), Machmouchi et al. (2017), Rosenfeld et al. (2017), Gruson et al. (2019), Schermann et al. (2016), Peska and Vojtas (2020), Chen et al. (2017b), Mattos et al. (2017b), Gupta and Chokshi (2020), Duivesteijn et al. (2017), Mattos et al. (2017a), Claeys et al. (2017), Tamburrelli and Margara (2014)
Concrete A/B testing problems	17	Saveski et al. (2017), Gui et al. (2015), Xu and Chen (2016), Li et al. (2021a), Spang et al. (2021), Dobaj et al. (2022), Wang et al. (2022), Bakshy and Frachtenberg (2015), Li et al. (2022), Liu et al. (2021), Goswami et al. (2015), Cai et al. (2021), Toslali et al. (2020), Satyal et al. (2019), Koukouvis et al. (2016), Lettner et al. (2013), Cámara and Kobsa (2009)
Pitfalls and challenges of A/B testing	14	Kohavi and Longbotham (2011), Kohavi et al. (2012), Esteller-Cucala et al. (2019), Fabijan et al. (2019), Xie et al. (2021), Dmitriev et al. (2017), Xu et al. (2015), Mattos et al. (2020), Appiktala et al. (2017), Dmitriev et al. (2016), Liu et al. (2019), Shariat et al. (2017), Kohavi et al. (2014), Deng et al. (2017)
Experimentation frameworks and platforms	13	Siqueira and de Paula (2018), Tu et al. (2021), Li et al. (2019), Vanderdonckt et al. (2019), Li et al. (2021b), Amatriain (2013), Reza et al. (2021), Gupchup et al. (2018), Borodovsky and Rosset (2011), De Souza et al. (2022), Zhao et al. (2017), Yu et al. (2020), Tang et al. (2010)
A/B testing at scale	9	Kohavi et al. (2013), Wang and Zhang (2021), Fabijan et al. (2017b), Xia et al. (2019), Jiang et al. (2020), Vasthimal et al. (2019), Gupta et al. (2018), Fabijan et al. (2017a, 2018)

5.1.6. Experimentation frameworks and platforms

This topic covers papers that present an A/B testing platform (Li et al., 2019; Tang et al., 2010; Siqueira and de Paula, 2018), or a framework concerning aspects related to the A/B testing process such as a framework for detecting data loss in A/B tests (Gupchup et al., 2018), a framework for the design of A/B tests (De Souza et al., 2022), or a framework for personalization of A/B testing (Tu et al., 2021).

5.1.7. A/B testing at scale

Primary studies under this topic focus on conducting A/B testing at a large scale, e.g., considerations for conducting A/B testing at scale (Jiang et al., 2020; Vasthimal et al., 2019; Gupta et al., 2018), process models or guidelines for A/B testing at scale (Fabijan et al., 2017b; Xia et al., 2019), or concrete scalable solutions such as a scalable statistical method for measuring quantile treatment effects for performance metrics in A/B tests (Wang and Zhang, 2021).

Table 20
Environments and tools used for A/B testing.

Environment	Number of occurrences
In-house experimentation system	21
Research tool or prototype	13
Commercial A/B testing tool	10
Commercial non A/B testing tool	7
User survey	1

5.2. Environments and tools used for A/B testing

In addition to the research topics covered in the primary studies, we also analyze the environments and tools that were used to realize A/B testing, see Table 20.

The most commonly mentioned type of environment is in-house experimentation system for A/B testing (21 occurrences), for instance

dedicated environments developed by companies such as Microsoft (Li et al., 2021a), Google (Tang et al., 2010), eBay (Vasthimal et al., 2019), and Etsy (Ju et al., 2019). These environments broadly support executing A/B tests. Furthermore, some primary studies describe concrete features of the experimentation system to help design A/B tests, e.g. controlling for bias during the specification of A/B tests in Airbnb's Experimentation Reporting Framework (Lee and Shen, 2018). Next, we observe research tools and prototypes (13 occurrences). Examples include a tool to perform online cloud experimentation (Toslali et al., 2020), a research prototype for A/B testing implemented in NodeJS (Schermann et al., 2016), a tool for A/B testing with decision assistants (Koukouvis et al., 2016), and a tool that enables automatic execution of multiple A/B tests (Tamburrelli and Margara, 2014). The remaining environments we identified were commercial A/B testing tools (10 occurrences), e.g., Optimizely (Miikulainen et al., 2020), and Google Analytics (Brown et al., 2016); commercial tools not related to A/B testing (7 occurrences), e.g., Crazy egg (Brown et al., 2016), a heatmapping tool used to design A/B variants, and using Yahoo Gemini (advertisement platform) to test different advertising strategies (Madlberger and Jizdny, 2021); and a user survey (1 occurrence) to determine which A/B variants to test by conducting a preliminary survey.

5.3. Research opportunities and future research directions

From our study, we propose a number of potential future research directions in the field of A/B testing. Concretely, we provide three lines of research: research on further improving the general process of A/B testing, research on automating aspects of A/B testing, and research on the adoption of proposed statistical methods in A/B testing.

5.3.1. Improving the A/B testing process

One future direction relates to taking considerations when running many A/B tests at once (Tang et al., 2010). Plenty of studies cover this topic by e.g., discussing lessons learned in unexpected A/B test results that were caused by other A/B tests that were running in parallel (Esteller-Cucala et al., 2019), or manually checking for possible effects of running A/B tests by analyzing the deployed A/B tests in the system (Vasthimal et al., 2019). Yet, we did not encounter a study that puts forward a systematic approach to tackle this problem.

Another avenue for future research is about improving the sensitivity in A/B tests by, e.g., combining different sensitivity improvement techniques as pointed out by Drutsa et al. (2015a), enabling proactive prediction of user behavior in A/B tests based on historical data (Drutsa et al., 2017a), and a deeper study of A/B test estimators to achieve better sensitivity as mentioned by Poyarkov et al. (2016).

The last avenue for future research in improving the A/B testing process relates to providing further guidelines and designing principles for choosing and engineering A/B metrics. We highlight two primary studies that mention open problems related to this opportunity: Kharitonov et al. (2017) put forward learning sensitive combinations of A/B metrics as a general open problem, and Duan et al. (2021) discuss investigating dynamics between surrogate metrics and the actual underlying metric.

5.3.2. Automation

In an effort to establish continuous experimentation, multiple studies put forward steps companies can take to develop an experimentation culture, e.g. Fabijan et al. (2017b), Xu et al. (2015), Fabijan et al. (2021). In light of expanding this experimentation culture, (partial) automation of the A/B testing process is essential to enable and empower continuous experimentation (Chen et al., 2019; Goswami et al., 2015). Initial research on automation of steps in the A/B testing has been conducted, as for example presented by Tamburrelli and Margara (2014) and Mattos et al. (2017b), see Sections 4.5.2 and 5.1.3. Yet the present state of research in this topic suggests that further investigation and more in-depth solutions are necessary to fully exploit automated design and execution of A/B tests. Some recent studies have explored tackling aspects of automating A/B testing more thoroughly, e.g., exploring the use of a multi-objective evolutionary algorithm to automatically create and evaluate A/B testing variants that can handle A/B tests with multiple A/B metrics of interest (Wu et al., 2023b), and a framework that partially takes over the manual analysis of A/B test results to assist stakeholders in making launch decisions (Wu et al., 2023a). Additionally, a number of open problems still remain that could facilitate and enable automated experimentation, e.g., determining which A/B tests to prioritize at execution (Goswami et al., 2015), and automatically generating insights related to the rationale and cause of experiment results to experiment developers to guide product development (Xu et al., 2015).

5.3.3. Adoption and tailoring statistical methods

Even though a number of primary studies discuss bootstrapping as a technique to evaluate the results of A/B tests (Tu et al., 2021; Agarwal et al., 2014; Goswami et al., 2015), bootstrapping remains largely unexplored in A/B testing, despite the fact that this statistical method has the potential to improve the analysis of A/B test results (Jiang et al., 2020; Bakshy and Frachtenberg, 2015). Moreover, bootstrapping can present an invaluable tool to provide statistical insights into the results of the tests which could e.g. not be obtained by a standard equality testing method (Efron and Tibshirani, 1994). However, one big downside of bootstrapping is that it is computationally expensive (Liu et al., 2019). Alongside adoption of known statistical methods, designing and tailoring new statistical methods to accommodate for particular experimentation scenarios presents an interesting research direction. One example is mentioned by Kharitonov et al. (2015), who put forward designing a custom statistical test for non-binomial A/B metrics. Another example concerns taking into account "the effects from multiple treatments with various metrics of interest" to tailor the approach presented by Tu et al. (2021) for optimal treatment assignments in A/B testing by leveraging causal effect estimations.

Besides a limited number of primary studies employing bootstrapping in the analysis of A/B tests, a significant number of studies mention statistically significant results or p-values in the analysis of conducted A/B tests without specifying the concrete statistical test used (37 occurrences). Moreover, a considerable number of studies do not report anything related to statistical analysis (47 occurrences). We argue that this information is important to report in research publications, and urge authors to specify the concrete statistical methods used[14] to obtain the results in the studies.

5.3.4. Exploration of the potential of recent AI developments

An interesting opportunity for future research testing is the use of recent developments in AI, such as Transformer and Large Language Models (LLM) (Vaswani et al., 2023; Wei et al., 2022; Kaplan et al., 2020). Tsigkanos et al. (2023) have already demonstrated the value of how a transformer-based LLM in the automation process of testing scientific software. Potential benefits of recent AI techniques could be enhancing the performance of A/B testing. For instance e.g., Stewart et al. (2023) propose an approach to support the classification of operational modes (or failure modes) of a system that could be used to improve the A/B testing process. Zhu et al. (2023) store experiences of system executions in memory or skill pool. Such experiences could be exploited to enhance the A/B process when the system encounters similar situations in the future. Nevertheless, several issues need further study before these AI techniques become practical, including the issue of hallucination (a model generates misleading and factually incorrect information), slow inference speed (due to the model size), explainability (such models have complex neural network architectures often operating as black boxes), and privacy concerns (models may share user data that may not align with users' data privacy expectations).

[14] Or alternatively an explicit mention of lack of statistical methods used.

5.4. Threats to validity

In this section we list potential threats to the validity of the systematic literature review (Ampatzoglou et al., 2019).

5.4.1. Internal validity

Internal validity refers to the extent to which a causal conclusion based on a study is warranted. One threat to the internal validity is a potential bias of researchers that perform the SLR, which may have an effect on the data collection and the insights derived in the study. In order to mitigate this threat, we involved multiple researchers in the study. Multiple researchers were responsible for selecting papers, extracting data and analyzing results. In each step, cross-checking was applied to minimize bias. Extra researchers were involved if no consensus could be found. Additionally, we defined a rigid protocol for the systematic literature review.

5.4.2. External validity

External validity refers to the extent to which the findings of the study can be generalized to the general field of A/B testing. A threat to the external validity of this systematic literature review is that some relevant primary studies may be not have been included. While we applied the search query to four main digital libraries that publish work in computer science, we have not included the Springer library. We excluded Springer since their search engine does not support querying abstracts of publications. Lack of an automatic process to retrieve studies would affect reproducability of the literature review. Yet, we acknowledge that we may have missed primary studies by excluding the Springer library in the search. To minimize this validity treat, we applied snowballing on the studies retrieved by the automatic search query on the four main libraries in order to uncover additional works that might have been missed.

5.4.3. Conclusion validity

Conclusion validity refers to the extent to which we obtained the right measure and whether we defined the right scope in relation to what is considered research in the field of A/B testing. One threat to the conclusion validity is the quality of the selected studies; studies of lower quality might produce insights that are not justified or applicable to the general field of A/B testing. In order to mitigate this threat, we excluded short papers, demo papers, and roadmap papers from the study. Furthermore, we evaluated a quality score for each selected paper. Papers with a quality score ≤ 4 were excluded from the study.

5.4.4. Reliability

Reliability refers to the extent to which this work is reproducible if the study would be conducted again. To mitigate this threat, we make all the collected and processed data available online. We also defined a specific search string, a list of online sources, and other specific details in the research protocol to ensure reproducibility. Bias of researchers also poses a threat here, influencing that similar results would be retrieved if the systematic literature review would be conducted again with a different set of reviewers.

6. Conclusion

A/B testing supports data-driven decisions about the adoption of features. It is widely used across different industries and key technology companies such as Google, Meta, and Microsoft. In this systematic literature review, we identified the subjects of A/B tests, how A/B tests are designed and executed, and the reported open research problems in the literature. We observed that algorithms, visual elements, and changes to a workflow or process are most commonly tested, with web, search engine, and e-commerce being the most popular application domains for A/B testing. Concerning the design of A/B tests, classic A/B tests with two variants are most commonly used, alongside engagement metrics such as conversion rate or number of impressions as metric to gauge the potential of the A/B variants. Hypothesis tests for equality testing are broadly utilized to analyze A/B test results, and bootstrapping also garners interest in a few primary studies. We devised three roles stakeholders take on in the design of A/B tests: Concept designer, Experiment architect, and Setup technician. Regarding the execution of A/B tests, empirical evaluation is the leading evaluation method. Besides the main A/B metrics, data concerning the product or system, and user-centric data are collected the most to conduct deeper analysis of the results of the A/B tests. A/B testing is most commonly used to determine and deploy the better performing A/B variant, or to gradually roll out a feature. Lastly, we devised two roles stakeholders take on in the execution of A/B tests: Experiment contributor, and Experiment assessor.

We identified seven categories of open problems: improving proposed approaches, extending the evaluation of the proposed approach, providing thorough analysis of the proposed approach, adding A/B testing process guidelines, automating the A/B testing process, enhancing scalability, and enhancing applicability. Leveraging these categories and observations made during the analysis, we provide three main lines of interesting research opportunities: developing more in-depth solutions to automate stages of the A/B testing process; presenting improvements to the A/B testing process by examining promising avenues for sensitivity improvement, systematic solutions to deal with interference of many A/B tests running at once, and providing guidelines and designing principles to choose and engineer A/B metrics; and lastly the adoption and tailoring of more sophisticated statistical methods such as bootstrapping to strengthen the analysis of A/B testing further. Finally, an interesting opportunity for future research would be to compare existing studies on A/B testing in terms of advantages and tradeoffs.

CRediT authorship contribution statement

Federico Quin: Conceptualization, Data curation, Methodology, Project administration, Software, Visualization, Writing – original draft, Writing – review & editing. **Danny Weyns:** Conceptualization, Methodology, Project administration, Supervision, Writing – review & editing. **Matthias Galster:** Methodology, Supervision, Writing – review & editing. **Camila Costa Silva:** Data curation, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgment

We thank Michiel Provoost for his support to this study.

Appendix. List of primary studies

See Table A.21.

Table A.21
List of primary studies.

Paper ID	Reference	Title
1	Abhishek and Mannor (2017)	A Nonparametric Sequential Test for Online Randomized Experiments
2	Saveski et al. (2017)	Detecting Network Effects: Randomizing Over Randomized Experiments
3	Kohavi and Longbotham (2011)	Unexpected Results in Online Controlled Experiments
4	Chen et al. (2019)	How A/B Tests Could Go Wrong: Automatic Diagnosis of Invalid Online Experiments
5	Poyarkov et al. (2016)	Boosted Decision Tree Regression Adjustment for Variance Reduction in Online Controlled Experiments
6	Kohavi et al. (2012)	Trustworthy Online Controlled Experiments: Five Puzzling Outcomes Explained
7	Kohavi et al. (2013)	Online Controlled Experiments at Large Scale
8	Wu et al. (2022)	Non-Stationary A/B Tests
9	Gui et al. (2015)	Network A/B Testing: From Sampling to Estimation
10	Xie et al. (2018)	False Discovery Rate Controlled Heterogeneous Treatment Effect Detection for Online Controlled Experiments
11	Esteller-Cucala et al. (2019)	Experimentation Pitfalls to Avoid in A/B Testing for Online Personalization
12	Deng et al. (2014)	Statistical Inference in Two-Stage Online Controlled Experiments with Treatment Selection and Validation
13	Budylin et al. (2018)	Consistent Transformation of Ratio Metrics for Efficient Online Controlled Experiments
14	Wang and Zhang (2021)	CONQ: CONtinuous Quantile Treatment Effects for Large-Scale Online Controlled Experiments
15	Fabijan et al. (2019)	Diagnosing Sample Ratio Mismatch in Online Controlled Experiments: A Taxonomy and Rules of Thumb for Practitioners
16	Siqueira and de Paula (2018)	IPEAD A/B Test Execution Framework
17	Johari et al. (2017)	Peeking at A/B Tests: Why It Matters, and What to Do about It
18	Kharitonov et al. (2017)	Learning Sensitive Combinations of A/B Test Metrics
19	Drutsa et al. (2015b)	Practical Aspects of Sensitivity in Online Experimentation with User Engagement Metrics
20	Xu and Chen (2016)	Evaluating Mobile Apps with A/B and Quasi A/B Tests
21	Deng et al. (2021)	On Post-Selection Inference in A/B Testing
22	Duan et al. (2021)	Online Experimentation with Surrogate Metrics: Guidelines and a Case Study
23	Drutsa et al. (2015a)	Future User Engagement Prediction and Its Application to Improve the Sensitivity of Online Experiments
24	Fabijan et al. (2017b)	The Evolution of Continuous Experimentation in Software Product Development: From Data to a Data-Driven Organization at Scale
25	Xie et al. (2021)	How to Measure Your App: A Couple of Pitfalls and Remedies in Measuring App Performance in Online Controlled Experiments
26	Kharitonov et al. (2015)	Sequential Testing for Early Stopping of Online Experiments
27	Dimmery et al. (2019)	Shrinkage Estimators in Online Experiments
28	Ha-Thuc et al. (2020)	A Counterfactual Framework for Seller-Side A/B Testing on Marketplaces
29	Drutsa et al. (2017a)	Periodicity in User Engagement with a Search Engine and Its Application to Online Controlled Experiments
30	Li et al. (2021a)	Evolving Software to be ML-Driven Utilizing Real-World A/B Testing: Experiences, Insights, Challenges
31	Drutsa et al. (2017b)	Using the Delay in a Treatment Effect to Improve Sensitivity and Preserve Directionality of Engagement Metrics in A/B Experiments
32	He et al. (2022)	A Cluster-Based Nearest Neighbor Matching Algorithm for Enhanced A/A Validation in Online Experimentation

(*continued on next page*)

Table A.21 (*continued*).

Paper ID	Reference	Title
33	Liou and Taylor (2020)	Variance-Weighted Estimators to Improve Sensitivity in Online Experiments
34	Dmitriev et al. (2017)	A Dirty Dozen: Twelve Common Metric Interpretation Pitfalls in Online Controlled Experiments
35	Lee and Shen (2018)	Winner's Curse: Bias Estimation for Total Effects of Features in Online Controlled Experiments
36	Xu et al. (2015)	From Infrastructure to Culture: A/B Testing Challenges in Large Scale Social Networks
37	Ju et al. (2019)	A Sequential Test for Selecting the Better Variant: Online A/B Testing, Adaptive Allocation, and Continuous Monitoring
38	Spang et al. (2021)	Unbiased Experiments in Congested Networks
39	Tu et al. (2021)	Personalized Treatment Selection Using Causal Heterogeneity
40	Machmouchi et al. (2017)	Beyond Success Rate: Utility as a Search Quality Metric for Online Experiments
41	Yoneda et al. (2019)	Algorithms and System Architecture for Immediate Personalized News Recommendations
42	Li et al. (2019)	Experimentation in the Operating System: The Windows Experimentation Platform
43	Vanderdonckt et al. (2019)	AB4Web: An On-Line A/B Tester for Comparing User Interface Design Alternatives
44	Okoshi et al. (2019)	Real-World Product Deployment of Adaptive Push Notification Scheduling on Smartphones
45	Kornilova and Bernardi (2021)	Mining the Stars: Learning Quality Ratings with User-Facing Explanations for Vacation Rentals
46	Dobaj et al. (2022)	Towards Digital Twin-Enabled DevOps for CPS Providing Architecture-Based Service Adaptation & Verification at Runtime
47	Wang et al. (2022)	Adaptive Experimentation with Delayed Binary Feedback
48	Li et al. (2021b)	Unifying Offline Causal Inference and Online Bandit Learning for Data Driven Decision
49	Amatriain (2013)	Beyond Data: From User Information to Business Value through Personalized Recommendations and Consumer Science
50	Li et al. (2021c)	Learning to Bundle Proactively for On-Demand Meal Delivery
51	Barajas et al. (2012b)	Measuring Dynamic Effects of Display Advertising in the Absence of User Tracking Information
52	Barajas et al. (2012a)	Marketing Campaign Evaluation in Targeted Display Advertising
53	Metrikov et al. (2014)	Whole Page Optimization: How Page Elements Interact with the Position Auction
54	Reza et al. (2021)	The MOOClet Framework: Unifying Experimentation, Dynamic Improvement, and Personalization in Online Courses
55	Xu et al. (2021)	Split-Treatment Analysis to Rank Heterogeneous Causal Effects for Prospective Interventions
56	Lalmas et al. (2015)	Promoting Positive Post-Click Experience for In-Stream Yahoo Gemini Users
57	Rosenfeld et al. (2017)	Predicting Counterfactuals from Large Historical Data and Small Randomized Trials
58	Sun et al. (2018)	Multi-Source Pointer Network for Product Title Summarization
59	Dai et al. (2021)	Beyond Relevance Ranking: A General Graph Matching Framework for Utility-Oriented Learning to Rank
60	Gruson et al. (2019)	Offline Evaluation to Make Decisions About PlaylistRecommendation Algorithms
61	Gupchup et al. (2018)	Trustworthy Experimentation Under Telemetry Loss
62	Gomez-Uribe and Hunt (2016)	The Netflix Recommender System: Algorithms, Business Value, and Innovation
63	Schermann et al. (2016)	Bifrost: Supporting Continuous Deployment with Automated Enactment of Multi-Phase Live Testing Strategies
64	Fu et al. (2020)	CompactETA: A Fast Inference System for Travel Time Prediction
65	Bakshy and Frachtenberg (2015)	Design and Analysis of Benchmarking Experiments for Distributed Internet Services
66	Ermis et al. (2020)	Learning to Rank in the Position Based Model with Bandit Feedback
67	Borisyuk et al. (2021)	VisRel: Media Search at Scale
68	Ye et al. (2022)	Behavioral Consequences of Reminder Emails on Students' Academic Performance: A Real-World Deployment
69	Li et al. (2017)	Content Recommendation by Noise Contrastive Transfer Learning of Feature Representation
70	Fairstein et al. (2022)	External Evaluation of Ranking Models under Extreme Position-Bias

(*continued on next page*)

Table A.21 (*continued*).

Paper ID	Reference	Title
71	Ueoka et al. (2020)	Tackling Cannibalization Problems for Online Advertisement
72	Tagami et al. (2014)	Filling Context-Ad Vocabulary Gaps with Click Logs
73	Freno (2017)	Practical Lessons from Developing a Large-Scale Recommender System at Zalando
74	Wu and Grbovic (2020)	How Airbnb Tells You Will Enjoy Sunset Sailing in Barcelona? Recommendation in a Two-Sided Travel Marketplace
75	Xu et al. (2014)	Modeling Professional Similarity by Mining Professional Career Trajectories
76	Chen et al. (2017a)	Social Incentive Optimization in Online Social Networks
77	Silberstein et al. (2020)	Ad Close Mitigation for Improved User Experience in Native Advertisements
78	Peska and Vojtas (2020)	Off-Line vs. On-Line Evaluation of Recommender Systems in Small E-Commerce
79	Agarwal et al. (2014)	LASER: A Scalable Response Prediction Platform for Online Advertising
80	Aiello et al. (2016)	The Role of Relevance in Sponsored Search
81	Sajeev et al. (2021)	Contextual Bandit Applications in a Customer Support Bot
82	Xia et al. (2019)	Safe Velocity: A Practical Guide to Software Deployment at Scale using Controlled Rollout
83	Szpektor et al. (2013)	When Relevance is Not Enough: Promoting Diversity and Freshness in Personalized Question Recommendation
84	Li et al. (2022)	Interference, Bias, and Variance in Two-Sided Marketplace Experimentation: Guidance for Platforms
85	Mattos et al. (2020)	Automotive A/B testing: Challenges and Lessons Learned from Practice
86	Borodovsky and Rosset (2011)	A/B Testing at SweetIM: The Importance of Proper Statistical Analysis
87	De Souza et al. (2022)	A Framework Model to Support A/B Tests at the Class and Component Level
88	Jiang et al. (2020)	Statistical Reasoning of Zero-Inflated Right-Skewed User-Generated Big Data A/B Testing
89	Liu et al. (2021)	Size matters? Or not: A/B testing with limited sample in automotive embedded software
90	Vasthimal et al. (2019)	Scalable Data Reporting Platform for A/B Tests
91	Blask (2013)	Applying Bayesian parameter estimation to A/B tests in e-business applications examining the impact of green marketing signals in sponsored search advertising
92	Alfaro-Flores et al. (2021)	Experiment-driven improvements in Human-in-the-loop Machine Learning Annotation via significance-based A/B testing
93	Gupta et al. (2018)	The Anatomy of a Large-Scale Experimentation Platform
94	Appiktala et al. (2017)	Demystifying dark matter for online experimentation
95	Goswami et al. (2015)	Controlled experiments for decision-making in e-Commerce search
96	Kumar and Hasteer (2017)	Evaluating usability of a web application: A comparative analysis of open-source tools
97	Chen et al. (2017b)	Faster online experimentation by eliminating traditional A/A validation
98	Dmitriev et al. (2016)	Pitfalls of long-term online controlled experiments
99	Sriraman et al. (2019)	SoftSKU: Optimizing Server Architectures for Microservice Diversity @Scale
100	Fabijan et al. (2017a)	The Benefits of Controlled Experimentation at Scale
101	Shi et al. (2015)	Context Adaptation for Smart Recommender Systems
102	Yang et al. (2018)	Whales, Dolphins, or Minnows? Towards the Player Clustering in Free Online Games Based on Purchasing Behavior via Data Mining Technique
103	Liu et al. (2019)	Enterprise-Level Controlled Experiments at Scale: Challenges and Solutions
104	Blask et al. (2011)	Should companies bid on their own brand in sponsored search?

(*continued on next page*)

Table A.21 (*continued*).

Paper ID	Reference	Title
105	He and Chen (2017)	A Probabilistic, Mechanism-Indepedent Outlier Detection Method for Online Experimentation
106	Zhao et al. (2017)	Inform Product Change through Experimentation with Data-Driven Behavioral Segmentation
107	Mattos et al. (2017b)	Your System Gets Better Every Day You Use It: Towards Automated Continuous Experimentation
108	Chakraborty et al. (2021)	Fashion Recommendation Systems, Models and Methods: A Review
109	Almeida and Casais (2022)	Subject Line Personalization Techniques and Their Influence in the E-Mail Marketing Open Rate
110	Madlberger and Jizdny (2021)	Impact of promotional social media content on click-through rate - Evidence from a FMCG company
111	Akase et al. (2021)	Related Entity Expansion and Ranking Using Knowledge Graph
112	Cai et al. (2021)	LinkLouvain: Link-Aware A/B Testing and Its Application on Online Marketing Campaign
113	Miikulainen et al. (2020)	Ascend by Evolv: Artificial intelligence-based massively multivariate conversion rate optimization
114	Yu et al. (2020)	A new framework for online testing of heterogeneous treatment effect
115	Toslali et al. (2020)	JACKPOT: Online experimentation of cloud microservices
116	Gupta and Chokshi (2020)	Digital Marketing Effectiveness Using Incrementality
117	Satyal et al. (2019)	Business process improvement with the AB-BPM methodology
118	Eckroth and Schoen (2019)	A genetic algorithm for finding a small and diverse set of recent news stories on a given subject: How we generate aaai's ai-alert
119	Kumar and Hosanagar (2017)	Measuring the value of recommendation links on product demand
120	Fabijan et al. (2018)	Experimentation growth: Evolving trustworthy A/B testing capabilities in online software companies
121	Shariat et al. (2017)	Online Evaluation of Bid Prediction Models in a Large-Scale Computational Advertising Platform: Decision Making and Insights
122	Satyal et al. (2017)	AB-BPM: Performance-driven instance routing for business process improvement
123	Duivesteijn et al. (2017)	Have It Both Ways—From A/B Testing to A&B Testing with Exceptional Model Mining
124	Mattos et al. (2017a)	More for Less: Automated Experimentation in Software-Intensive Systems
125	Claeys et al. (2017)	Regression Tree for Bandits Models in A/B Testing
126	Pelleg et al. (2016)	When the Crowd is Not Enough: Improving User Experience with Social Media through Automatic Quality Analysis
127	Brown et al. (2016)	Pixel efficiency analysis: A quantitative web analytics approach
128	Koukouvis et al. (2016)	A/B Testing in E-commerce Sales Processes
129	Ozawa et al. (2016)	A Method for the Construction of User Targeting Knowledge for B2B Industry Website
130	Qu and Zhang (2016)	Validating Mobile Designs with Agile Testing in China: Based on Baidu Map for Mobile
131	Wang and Hardtke (2015)	User Latent Preference Model for Better Downside Management in Recommender Systems
132	Tamburrelli and Margara (2014)	Towards Automated A/B Testing
133	Kohavi et al. (2014)	Seven rules of thumb for web site experimenters
134	Lettner et al. (2013)	Enabling A/B Testing of Native Mobile Applications by Remote User Interface Exchange
135	Tang et al. (2010)	Overlapping Experiment Infrastructure: More, Better, Faster Experimentation
136	Funk (2009)	Optimizing price levels in e-commerce applications: An empirical study

(*continued on next page*)

Table A.21 (*continued*).

Paper ID	Reference	Title
137	Cámara and Kobsa (2009)	Facilitating Controlled Tests of Website Design Changes: A Systematic Approach
138	Aharon et al. (2019a)	Soft Frequency Capping for Improved Ad Click Prediction in Yahoo Gemini Native
139	Deng (2015)	Objective Bayesian Two Sample Hypothesis Testing for Online Controlled Experiments
140	Feit and Berman (2019)	Test & Roll: Profit-Maximizing A/B Tests
141	Young (2014)	Improving Library User Experience with A/B Testing: Principles and Process

References

Abhishek, Vineet, Mannor, Shie, 2017. A nonparametric sequential test for online randomized experiments. In: Proceedings of the 26th International Conference on World Wide Web Companion. In: WWW '17 Companion, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, pp. 610–616. http://dx.doi.org/10.1145/3041021.3054196.

Agarwal, Deepak, Long, Bo, Traupman, Jonathan, Xin, Doris, Zhang, Liang, 2014. LASER: A scalable response prediction platform for online advertising. In: Proceedings of the 7th ACM International Conference on Web Search and Data Mining. WSDM '14, Association for Computing Machinery, New York, NY, USA, pp. 173–182. http://dx.doi.org/10.1145/2556195.2556252.

Aharon, Michal, Kaplan, Yohay, Levy, Rina, Somekh, Oren, Blanc, Ayelet, Eshel, Neetai, Shahar, Avi, Singer, Assaf, Zlotnik, Alex, 2019a. Soft frequency capping for improved ad click prediction in yahoo gemini native. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. CIKM '19, Association for Computing Machinery, New York, NY, USA, pp. 2793–2801. http://dx.doi.org/10.1145/3357384.3357801.

Aharon, Michal, Somekh, Oren, Shahar, Avi, Singer, Assaf, Trayvas, Baruch, Vogel, Hadas, Dobrev, Dobri, 2019b. Carousel ads optimization in yahoo gemini native. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. KDD '19, Association for Computing Machinery, New York, NY, USA, pp. 1993–2001. http://dx.doi.org/10.1145/3292500.3330740.

Aiello, Luca, Arapakis, Ioannis, Baeza-Yates, Ricardo, Bai, Xiao, Barbieri, Nicola, Mantrach, Amin, Silvestri, Fabrizio, 2016. The role of relevance in sponsored search. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. CIKM '16, Association for Computing Machinery, New York, NY, USA, pp. 185–194. http://dx.doi.org/10.1145/2983323.2983840.

Akase, Ryuya, Kawabata, Hiroto, Nishida, Akiomi, Tanaka, Yuki, Kaminaga, Tamaki, 2021. Related entity expansion and ranking using knowledge graph. In: Barolli, Leonard, Yim, Kangbin, Enokido, Tomoya (Eds.), Complex, Intelligent and Software Intensive Systems. Springer International Publishing, Cham, pp. 172–184.

Alfaro-Flores, Rafael, Salas-Bonilla, José, Juillard, Loic, Esquivel-Rodríguez, Juan, 2021. Experiment-driven improvements in human-in-the-loop machine learning annotation via significance-based A/B testing. In: 2021 XLVII Latin American Computing Conference. CLEI, pp. 1–9. http://dx.doi.org/10.1109/CLEI53233.2021.9639977.

Almeida, Joana, Casais, Beatriz, 2022. Subject line personalization techniques and their influence in the E-mail marketing open rate. In: Rocha, Alvaro, Adeli, Hojjat, Dzemyda, Gintautas, Moreira, Fernando (Eds.), Information Systems and Technologies. Springer International Publishing, Cham, pp. 532–540.

Amatriain, Xavier, 2013. Beyond data: From user information to business value through personalized recommendations and consumer science. In: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. CIKM '13, Association for Computing Machinery, New York, NY, USA, pp. 2201–2208. http://dx.doi.org/10.1145/2505515.2514701.

Ampatzoglou, Apostolos, Bibi, Stamatia, Avgeriou, Paris, Verbeek, Marijn, Chatzigeorgiou, Alexander, 2019. Identifying, categorizing and mitigating threats to validity in software engineering secondary studies. Inf. Softw. Technol. 106, 201–230. http://dx.doi.org/10.1016/j.infsof.2018.10.006.

Appiktala, Nirupama, Chen, Miao, Natkovich, Michael, Walters, Joshua, 2017. Demystifying dark matter for online experimentation. In: 2017 IEEE International Conference on Big Data (Big Data). pp. 1620–1626. http://dx.doi.org/10.1109/BigData.2017.8258096.

Auer, F., Felderer, M., 2018. Current state of research on continuous experimentation: A systematic mapping study. In: 2018 44th Euromicro Conference on Software Engineering and Advanced Applications. SEAA, IEEE Computer Society, Los Alamitos, CA, USA, pp. 335–344. http://dx.doi.org/10.1109/SEAA.2018.00062.

Auer, Florian, Ros, Rasmus, Kaltenbrunner, Lukas, Runeson, Per, Felderer, Michael, 2021. Controlled experimentation in continuous experimentation: Knowledge and challenges. Inf. Softw. Technol. 134, 106551. http://dx.doi.org/10.1016/j.infsof.2021.106551.

Bakshy, Eytan, Frachtenberg, Eitan, 2015. Design and analysis of benchmarking experiments for distributed internet services. In: Proceedings of the 24th International Conference on World Wide Web. WWW '15, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, pp. 108–118. http://dx.doi.org/10.1145/2736277.2741082.

Barajas, Joel, Kwon, Jaimie, Akella, Ram, Flores, Aaron, Holtan, Marius, Andrei, Victor, 2012a. Marketing campaign evaluation in targeted display advertising. In: Proceedings of the Sixth International Workshop on Data Mining for Online Advertising and Internet Economy. ADKDD '12, Association for Computing Machinery, New York, NY, USA, http://dx.doi.org/10.1145/2351356.2351361.

Barajas, Joel, Kwon, Jaimie, Akella, Ram, Flores, Aaron, Holtan, Marius, Andrei, Victor, 2012b. Measuring dynamic effects of display advertising in the absence of user tracking information. In: Proceedings of the Sixth International Workshop on Data Mining for Online Advertising and Internet Economy. ADKDD '12, Association for Computing Machinery, New York, NY, USA, http://dx.doi.org/10.1145/2351356.2351364.

Basili, Victor R., Caldiera, Gianluigi, Rombach, Dieter H., 1994. The goal question metric approach. I, John Wiley & Sons,

Blask, Tobias, 2013. Applying Bayesian parameter estimation to A/B tests in e-business applications examining the impact of green marketing signals in sponsored search advertising. In: 2013 International Conference on E-Business (ICE-B). pp. 1–8.

Blask, Tobias, Funk, Burkhardt, Schulte, Reinhard, 2011. Should companies bid on their own brand in sponsored search? In: Proceedings of the International Conference on E-Business. pp. 1–8.

Borisyuk, Fedor, Malreddy, Siddarth, Mei, Jun, Liu, Yiqun, Liu, Xiaoyi, Maheshwari, Piyush, Bell, Anthony, Rangadurai, Kaushik, 2021. VisRel: Media search at scale. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. KDD '21, Association for Computing Machinery, New York, NY, USA, pp. 2584–2592. http://dx.doi.org/10.1145/3447548.3467081.

Borodovsky, Slava, Rosset, Saharon, 2011. A/b testing at sweetim: The importance of proper statistical analysis. In: 2011 IEEE 11th International Conference on Data Mining Workshops. pp. 733–740. http://dx.doi.org/10.1109/ICDMW.2011.19.

Brown, Alex, Lush, Binky, Jansen, Bernard J., 2016. Pixel efficiency analysis: A quantitative web analytics approach. In: Proceedings of the Association for Information Science and Technology. Vol. 53, (1), pp. 1–10. http://dx.doi.org/10.1002/pra2.2016.14505301040, arXiv:https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/pra2.2016.14505301040.

Budylin, Roman, Drutsa, Alexey, Katsev, Ilya, Tsoy, Valeriya, 2018. Consistent transformation of ratio metrics for efficient online controlled experiments. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. WSDM '18, Association for Computing Machinery, New York, NY, USA, pp. 55–63. http://dx.doi.org/10.1145/3159652.3159699.

Cai, Tianchi, Cheng, Daxi, Liang, Chen, Liu, Ziqi, Gu, Lihong, Xie, Huizhi, Zhang, Zhiqiang, Zeng, Xiaodong, Gu, Jinjie, 2021. LinkLouvain: Link-aware A/B testing and its application on online marketing campaign. In: Jensen, Christian S., Lim, Ee-Peng, Yang, De-Nian, Lee, Wang-Chien, Tseng, Vincent S., Kalogeraki, Vana, Huang, Jen-Wei, Shen, Chih-Ya (Eds.), Database Systems for Advanced Applications. Springer International Publishing, Cham, pp. 499–510.

Cámara, Javier, Kobsa, Alfred, 2009. Facilitating controlled tests of website design changes: A systematic approach. In: Gaedke, Martin, Grossniklaus, Michael, Díaz, Oscar (Eds.), Web Engineering. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 370–378.

Chakraborty, Samit, Hoque, Md. Saiful, Rahman Jeem, Naimur, Biswas, Manik Chandra, Bardhan, Deepayan, Lobaton, Edgar, 2021. Fashion recommendation systems, models and methods: A review. Informatics 8 (3), http://dx.doi.org/10.3390/informatics8030049, URL https://www.mdpi.com/2227-9709/8/3/49.

Chen, Guangde, Chen, Bee-Chung, Agarwal, Deepak, 2017a. Social incentive optimization in online social networks. In: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. WSDM '17, Association for Computing Machinery, New York, NY, USA, pp. 547–556. http://dx.doi.org/10.1145/3018661.3018700.

Chen, Russell, Chen, Miao, Jadav, Mahendrasinh Ramsinh, Bae, Joonsuk, Matheson, Don, 2017b. Faster online experimentation by eliminating traditional A/A validation. In: 2017 IEEE International Conference on Big Data (Big Data). pp. 1635–1641. http://dx.doi.org/10.1109/BigData.2017.8258098.

Chen, Nanyu, Liu, Min, Xu, Ya, 2019. How A/B tests could go wrong: Automatic diagnosis of invalid online experiments. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. WSDM '19, Association for Computing Machinery, New York, NY, USA, pp. 501–509. http://dx.doi.org/10.1145/3289600.3291000.

Claeys, Emmanuelle, Gançarski, Pierre, Maumy-Bertrand, Myriam, Wassner, Hubert, 2017. Regression tree for bandits models in A/B testing. In: Adams, Niall, Tucker, Allan, Weston, David (Eds.), Advances in Intelligent Data Analysis XVI. Springer International Publishing, Cham, pp. 52–62.

Costa, Rafael, Cheniaux, Elie, Rosaes, Pedro, Carvalho, Marcele, Freire, Rafael, Versiani, Márcio, Range, Bernard, Nardi, Antonio, 2011. The effectiveness of cognitive behavioral group therapy in treating bipolar disorder: A randomized controlled study. Revista brasileira de psiquiatria (São Paulo, Brazil : 1999) 33, 144–149. http://dx.doi.org/10.1590/S1516-44462011000200009.

Creswell, John, Guetterman, Timothy, 2018. Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research, sixth ed. Pearson, New York, NY, USA.

Dai, Xinyi, Xi, Yunjia, Zhang, Weinan, Liu, Qing, Tang, Ruiming, He, Xiuqiang, Hou, Jiawei, Wang, Jun, Yu, Yong, 2021. Beyond relevance ranking: A general graph matching framework for utility-oriented learning to rank. ACM Trans. Inf. Syst. 40 (2), http://dx.doi.org/10.1145/3464303.

Daneva, Maya, Damian, Daniela, Marchetto, Alessandro, Pastor, Oscar, 2014. Empirical research methodologies and studies in requirements engineering: How far did we come? J. Syst. Softw. 95, 1–9. http://dx.doi.org/10.1016/j.jss.2014.06.035.

de Feijter, Rico, Overbeek, Sietse, van Vliet, Rob, Jagroep, Erik, Brinkkemper, Sjaak, 2018. DevOps competences and maturity for software producing organizations. In: Gulden, Jens, Reinhartz-Berger, Iris, Schmidt, Rainer, Guerreiro, Sérgio, Guédria, Wided, Bera, Palash (Eds.), Enterprise, Business-Process and Information Systems Modeling. Springer International Publishing, Cham, pp. 244–259.

De Souza, Wagner S., Pereira, Fernando O., Albuquerque, Vanessa G., Melegati, Jorge, Guerra, Eduardo, 2022. A framework model to support A/B tests at the class and component level. In: 2022 IEEE 46th Annual Computers, Software, and Applications Conference. COMPSAC, pp. 860–865. http://dx.doi.org/10.1109/COMPSAC54236.2022.00136.

Deng, Alex, 2015. Objective Bayesian two sample hypothesis testing for online controlled experiments. In: Proceedings of the 24th International Conference on World Wide Web. In: WWW '15 Companion, Association for Computing Machinery, New York, NY, USA, pp. 923–928. http://dx.doi.org/10.1145/2740908.2742563.

Deng, Alex, Li, Tianxi, Guo, Yu, 2014. Statistical inference in two-stage online controlled experiments with treatment selection and validation. In: Proceedings of the 23rd International Conference on World Wide Web. WWW '14, Association for Computing Machinery, New York, NY, USA, pp. 609–618. http://dx.doi.org/10.1145/2566486.2568028.

Deng, Alex, Li, Yicheng, Lu, Jiannan, Ramamurthy, Vivek, 2021. On post-selection inference in A/B testing. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. KDD '21, Association for Computing Machinery, New York, NY, USA, pp. 2743–2752. http://dx.doi.org/10.1145/3447548.3467129.

Deng, Alex, Lu, Jiannan, Litz, Jonthan, 2017. Trustworthy analysis of online A/B tests: Pitfalls, challenges and solutions. In: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. WSDM '17, Association for Computing Machinery, New York, NY, USA, pp. 641–649. http://dx.doi.org/10.1145/3018661.3018677.

Deng, Alex, Xu, Ya, Kohavi, Ron, Walker, Toby, 2013. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining. WSDM '13, Association for Computing Machinery, New York, NY, USA, pp. 123–132. http://dx.doi.org/10.1145/2433396.2433413.

Dimmery, Drew, Bakshy, Eytan, Sekhon, Jasjeet, 2019. Shrinkage estimators in online experiments. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. KDD '19, Association for Computing Machinery, New York, NY, USA, pp. 2914–2922. http://dx.doi.org/10.1145/3292500.3330771.

Dmitriev, Pavel, Frasca, Brian, Gupta, Somit, Kohavi, Ron, Vaz, Garnet, 2016. Pitfalls of long-term online controlled experiments. In: 2016 IEEE International Conference on Big Data (Big Data). pp. 1367–1376. http://dx.doi.org/10.1109/BigData.2016.7840744.

Dmitriev, Pavel, Gupta, Somit, Kim, Dong Woo, Vaz, Garnet, 2017. A dirty dozen: Twelve common metric interpretation pitfalls in online controlled experiments. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '17, Association for Computing Machinery, New York, NY, USA, pp. 1427–1436. http://dx.doi.org/10.1145/3097983.3098024.

Dobaj, Jürgen, Riel, Andreas, Krug, Thomas, Seidl, Matthias, Macher, Georg, Egretzberger, Markus, 2022. Towards digital twin-enabled DevOps for CPS providing architecture-based service adaptation & verification at runtime. In: Proceedings of the 17th Symposium on Software Engineering for Adaptive and Self-Managing Systems. SEAMS '22, Association for Computing Machinery, New York, NY, USA, pp. 132–143. http://dx.doi.org/10.1145/3524844.3528057.

Drutsa, Alexey, Gusev, Gleb, Serdyukov, Pavel, 2015a. Future user engagement prediction and its application to improve the sensitivity of online experiments. In: Proceedings of the 24th International Conference on World Wide Web. WWW '15, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, pp. 256–266. http://dx.doi.org/10.1145/2736277.2741116.

Drutsa, Alexey, Gusev, Gleb, Serdyukov, Pavel, 2017a. Periodicity in user engagement with a search engine and its application to online controlled experiments. ACM Trans. Web 11 (2), http://dx.doi.org/10.1145/2856822.

Drutsa, Alexey, Gusev, Gleb, Serdyukov, Pavel, 2017b. Using the delay in a treatment effect to improve sensitivity and preserve directionality of engagement metrics in A/B experiments. In: Proceedings of the 26th International Conference on World Wide Web. WWW '17, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, pp. 1301–1310. http://dx.doi.org/10.1145/3038912.3052664.

Drutsa, Alexey, Ufland, Anna, Gusev, Gleb, 2015b. Practical aspects of sensitivity in online experimentation with user engagement metrics. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. CIKM '15, Association for Computing Machinery, New York, NY, USA, pp. 763–772. http://dx.doi.org/10.1145/2806416.2806496.

Duan, Weitao, Ba, Shan, Zhang, Chunzhe, 2021. Online experimentation with surrogate metrics: Guidelines and a case study. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining. WSDM '21, Association for Computing Machinery, New York, NY, USA, pp. 193–201. http://dx.doi.org/10.1145/3437963.3441737.

Duivesteijn, Wouter, Farzami, Tara, Putman, Thijs, Peer, Evertjan, Weerts, Hilde J.P., Adegeest, Jasper N., Foks, Gerson, Pechenizkiy, Mykola, 2017. Have it both ways—From A/B testing to A&B testing with exceptional model mining. In: Altun, Yasemin, Das, Kamalika, Mielikäinen, Taneli, Malerba, Donato, Stefanowski, Jerzy, Read, Jesse, Žitnik, Marinka, Ceci, Michelangelo, Džeroski, Sašo (Eds.), Machine Learning and Knowledge Discovery in Databases. Springer International Publishing, Cham, pp. 114–126.

Eckroth, Joshua, Schoen, Eric, 2019. A genetic algorithm for finding a small and diverse set of recent news stories on a given subject: How we generate aaai's ai-alert. In: 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019. pp. 9357–9364, URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85090801224&partnerID=40&md5=f3391d595e00df8a0cba7802c9043ebd, Cited by: 2.

Efron, B., Tibshirani, R.J., 1994. An Introduction to the Bootstrap. CRC Press, URL https://books.google.be/books?id=MWC1DwAAQBAJ.

Ermis, Beyza, Ernst, Patrick, Stein, Yannik, Zappella, Giovanni, 2020. Learning to rank in the position based model with bandit feedback. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management. CIKM '20, Association for Computing Machinery, New York, NY, USA, pp. 2405–2412. http://dx.doi.org/10.1145/3340531.3412723.

Erthal, Vladimir M., de Souza, Bruno P., dos Santos, Paulo Sérgio M., Travassos, Guilherme H., 2022. A literature study to characterize continuous experimentation in software engineering. In: CIbSE 2022 - XXV Ibero-American Conference on Software Engineering. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85137064966&partnerID=40&md5=04240b73ab90eb841083173be558b33f, Cited by: 0.

Esteller-Cucala, Maria, Fernandez, Vicenc, Villuendas, Diego, 2019. Experimentation pitfalls to avoid in a/b testing for online personalization. In: Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization. In: UMAP'19 Adjunct, Association for Computing Machinery, New York, NY, USA, pp. 153–159. http://dx.doi.org/10.1145/3314183.3323853.

Fabijan, Aleksander, Arai, Benjamin, Dmitriev, Pavel, Vermeer, Lukas, 2021. It takes a flywheel to fly: Kickstarting and growing the a/b testing momentum at scale. In: 2021 47th Euromicro Conference on Software Engineering and Advanced Applications. SEAA, pp. 109–118. http://dx.doi.org/10.1109/SEAA53835.2021.00023.

Fabijan, Aleksander, Dmitriev, Pavel, McFarland, Colin, Vermeer, Lukas, Holmström Olsson, Helena, Bosch, Jan, 2018. Experimentation growth: Evolving trustworthy A/B testing capabilities in online software companies. J. Softw.: Evol. Process 30 (12), e2113. http://dx.doi.org/10.1002/smr.2113, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/smr.2113, e2113 JSME-17-0210.R2.

Fabijan, Aleksander, Dmitriev, Pavel, Olsson, Helena Holmström, Bosch, Jan, 2017a. The benefits of controlled experimentation at scale. In: 2017 43rd Euromicro Conference on Software Engineering and Advanced Applications. SEAA, pp. 18–26. http://dx.doi.org/10.1109/SEAA.2017.47.

Fabijan, Aleksander, Dmitriev, Pavel, Olsson, Helena Holmström, Bosch, Jan, 2017b. 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. ICSE '17, IEEE Press, Los Alamitos, CA, USA, pp. 770–780. http://dx.doi.org/10.1109/ICSE.2017.76.

Fabijan, A., Dmitriev, P., Olsson, H. Holmstrom, Bosch, J., 2020. The online controlled experiment lifecycle. IEEE Softw. 37 (02), 60–67. http://dx.doi.org/10.1109/MS.2018.2875842.

Fabijan, Aleksander, Gupchup, Jayant, Gupta, Somit, Omhover, Jeff, Qin, Wen, Vermeer, Lukas, Dmitriev, Pavel, 2019. Diagnosing sample ratio mismatch in online controlled experiments: A taxonomy and rules of thumb for practitioners. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. KDD '19, Association for Computing Machinery, New York, NY, USA, pp. 2156–2164. http://dx.doi.org/10.1145/3292500.3330722.

Fabijan, Aleksander, Olsson, Helena Holmström, Bosch, Jan, 2015. Customer feedback and data collection techniques in software R&D: A literature review. In: Fernandes, João M., Machado, Ricardo J., Wnuk, Krzysztof (Eds.), Software Business. Springer International Publishing, Cham, pp. 139–153. http://dx.doi.org/10.1007/978-3-319-19593-3_12.

Fabijan, Aleksander, Olsson, Helena Holmström, Bosch, Jan, 2016. The lack of sharing of customer data in large software organizations: Challenges and implications. In: Sharp, Helen, Hall, Tracy (Eds.), Agile Processes, in Software Engineering, and Extreme Programming. Springer International Publishing, Cham, pp. 39–52. http://dx.doi.org/10.1007/978-3-319-33515-5_4.

Fairstein, Yaron, Haramaty, Elad, Lazerson, Arnon, Lewin-Eytan, Liane, 2022. External evaluation of ranking models under extreme position-bias. In: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining. WSDM '22, Association for Computing Machinery, New York, NY, USA, pp. 252–261. http://dx.doi.org/10.1145/3488560.3498420.

Feit, Elea McDonnell, Berman, Ron, 2019. Test & roll: Profit-maximizing A/B tests. Mark. Sci. 38 (6), 1038–1058. http://dx.doi.org/10.1287/mksc.2019.1194.

Freno, Antonino, 2017. Practical lessons from developing a large-scale recommender system at zalando. In: Proceedings of the Eleventh ACM Conference on Recommender Systems. RecSys '17, Association for Computing Machinery, New York, NY, USA, pp. 251–259. http://dx.doi.org/10.1145/3109859.3109897.

Fu, Kun, Meng, Fanlin, Ye, Jieping, Wang, Zheng, 2020. Compacteta: A fast inference system for travel time prediction. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. KDD '20, Association for Computing Machinery, New York, NY, USA, pp. 3337–3345. http://dx.doi.org/10.1145/3394486.3403386.

Funk, Burkhardt, 2009. Optimizing price levels in e-commerce applications: An empirical study. In: ICETE 2009 - International Joint Conference on e-Business and Telecommunications. pp. 37–43, URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-74549181430&partnerID=40&md5=6dfdde67b807b3964c62fc8c1929dcf0, Cited by: 1.

Galster, Matthias, Weyns, Danny, 2016. Empirical research in software architecture: How far have we come? In: 2016 13th Working IEEE/IFIP Conference on Software Architecture. WICSA, IEEE Press, Los Alamitos, CA, USA, pp. 11–20. http://dx.doi.org/10.1109/WICSA.2016.10.

Giaimo, Federico, Andrade, Hugo, Berger, Christian, 2020. Continuous experimentation and the cyber–physical systems challenge: An overview of the literature and the industrial perspective. J. Syst. Softw. 170, 110781. http://dx.doi.org/10.1016/j.jss.2020.110781.

Gomez-Uribe, Carlos A., Hunt, Neil, 2016. The netflix recommender system: Algorithms, business value, and innovation. ACM Trans. Manage. Inf. Syst. 6 (4), http://dx.doi.org/10.1145/2843948.

Goswami, Anjan, Han, Wei, Wang, Zhenrui, Jiang, Angela, 2015. Controlled experiments for decision-making in e-commerce search. In: 2015 IEEE International Conference on Big Data (Big Data). IEEE Press, Los Alamitos, CA, USA, pp. 1094–1102. http://dx.doi.org/10.1109/BigData.2015.7363863.

Gruson, Alois, Chandar, Praveen, Charbuillet, Christophe, McInerney, James, Hansen, Samantha, Tardieu, Damien, Carterette, Ben, 2019. Offline evaluation to make decisions about PlaylistRecommendation algorithms. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. WSDM '19, Association for Computing Machinery, New York, NY, USA, pp. 420–428. http://dx.doi.org/10.1145/3289600.3291027.

Gui, Huan, Xu, Ya, Bhasin, Anmol, Han, Jiawei, 2015. Network A/B testing: From sampling to estimation. In: Proceedings of the 24th International Conference on World Wide Web. WWW '15, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, pp. 399–409. http://dx.doi.org/10.1145/2736277.2741081.

Gupchup, Jayant, Hosseinkashi, Yasaman, Dmitriev, Pavel, Schneider, Daniel, Cutler, Ross, Jefremov, Andrei, Ellis, Martin, 2018. Trustworthy experimentation under telemetry loss. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management. CIKM '18, Association for Computing Machinery, New York, NY, USA, pp. 387–396. http://dx.doi.org/10.1145/3269206.3271747.

Gupta, Shubham, Chokshi, Sneha, 2020. Digital marketing effectiveness using incrementality. In: Singh, Mayank, Gupta, P.K., Tyagi, Vipin, Flusser, Jan, Ören, Tuncer, Valentino, Gianluca (Eds.), Advances in Computing and Data Sciences. Springer Singapore, Singapore, pp. 66–75.

Gupta, Somit, Kohavi, Ronny, Tang, Diane, Xu, Ya, Andersen, Reid, Bakshy, Eytan, Cardin, Niall, Chandran, Sumita, Chen, Nanyu, Coey, Dominic, et al., 2019. Top challenges from the first practical online controlled experiments summit. ACM SIGKDD Explor. Newsl. 21 (1), 20–35.

Gupta, Somit, Ulanova, Lucy, Bhardwaj, Sumit, Dmitriev, Pavel, Raff, Paul, Fabijan, Aleksander, 2018. The anatomy of a large-scale experimentation platform. In: 2018 IEEE International Conference on Software Architecture. ICSA, pp. 1–109. http://dx.doi.org/10.1109/ICSA.2018.00009.

Ha-Thuc, Viet, Dutta, Avishek, Mao, Ren, Wood, Matthew, Liu, Yunli, 2020. A counterfactual framework for seller-side A/B testing on marketplaces. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '20, Association for Computing Machinery, New York, NY, USA, pp. 2288–2296. http://dx.doi.org/10.1145/3397271.3401434.

He, Yan, Chen, Miao, 2017. A probabilistic, mechanism-indepedent outlier detection method for online experimentation. In: 2017 IEEE International Conference on Data Science and Advanced Analytics. DSAA, pp. 640–647. http://dx.doi.org/10.1109/DSAA.2017.64.

He, Yan, Yu, Lin, Chen, Miao, Choi, William, Matheson, Don, 2022. A cluster-based nearest neighbor matching algorithm for enhanced a/a validation in online experimentation. In: Companion Proceedings of the Web Conference 2022. WWW '22, Association for Computing Machinery, New York, NY, USA, pp. 136–140. http://dx.doi.org/10.1145/3487553.3524220.

Humble, Jez, Farley, David, 2010. Continuous Delivery: Reliable Software Releases through Build, Test, and Deployment Automation, first ed. Addison-Wesley Professional, Illinois, IL, USA.

Jiang, Hao, Yang, Fan, Wei, Wutao, 2020. Statistical reasoning of zero-inflated right-skewed user-generated big data A/B testing. In: 2020 IEEE International Conference on Big Data (Big Data). pp. 1533–1544. http://dx.doi.org/10.1109/BigData50022.2020.9377996.

Johari, Ramesh, Koomen, Pete, Pekelis, Leonid, Walsh, David, 2017. Peeking at A/B tests: Why it matters, and what to do about it. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '17, Association for Computing Machinery, New York, NY, USA, pp. 1517–1525. http://dx.doi.org/10.1145/3097983.3097992.

Ju, Nianqiao, Hu, Diane, Henderson, Adam, Hong, Liangjie, 2019. A sequential test for selecting the better variant: Online A/B testing, adaptive allocation, and continuous monitoring. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. WSDM '19, Association for Computing Machinery, New York, NY, USA, pp. 492–500. http://dx.doi.org/10.1145/3289600.3291025.

Kaplan, Jared, McCandlish, Sam, Henighan, Tom, Brown, Tom B., Chess, Benjamin, Child, Rewon, Gray, Scott, Radford, Alec, Wu, Jeffrey, Amodei, Dario, 2020. Scaling laws for neural language models. arXiv:2001.08361.

Keele, Staffs, et al., 2007. Guidelines for performing systematic literature reviews in software engineering. Technical report, Technical report, Ver. 2.3 EBSE Technical Report. EBSE.

Kharitonov, Eugene, Drutsa, Alexey, Serdyukov, Pavel, 2017. Learning sensitive combinations of A/B test metrics. In: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. WSDM '17, Association for Computing Machinery, New York, NY, USA, pp. 651–659. http://dx.doi.org/10.1145/3018661.3018708.

Kharitonov, Eugene, Vorobev, Aleksandr, Macdonald, Craig, Serdyukov, Pavel, Ounis, Iadh, 2015. Sequential testing for early stopping of online experiments. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '15, Association for Computing Machinery, New York, NY, USA, pp. 473–482. http://dx.doi.org/10.1145/2766462.2767729.

King, Rochelle, Churchill, Elizabeth F., Tan, Caitlin, 2017. Designing with Data: Improving the User Experience with A/B Testing. O'Reilly Media, Inc., Sebastopol, CA, USA.

Kohavi, Ron, Deng, Alex, Frasca, Brian, Longbotham, Roger, Walker, Toby, Xu, Ya, 2012. Trustworthy online controlled experiments: Five puzzling outcomes explained. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '12, Association for Computing Machinery, New York, NY, USA, pp. 786–794. http://dx.doi.org/10.1145/2339530.2339653.

Kohavi, Ron, Deng, Alex, Frasca, Brian, Walker, Toby, Xu, Ya, Pohlmann, Nils, 2013. Online controlled experiments at large scale. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '13, Association for Computing Machinery, New York, NY, USA, pp. 1168–1176. http://dx.doi.org/10.1145/2487575.2488217.

Kohavi, Ron, Deng, Alex, Longbotham, Roger, Xu, Ya, 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, NY, USA, pp. 1857–1866. http://dx.doi.org/10.1145/2623330.2623341.

Kohavi, Ron, Longbotham, Roger, 2011. Unexpected results in online controlled experiments. SIGKDD Explor. Newsl. 12 (2), 31–35. http://dx.doi.org/10.1145/1964897.1964905.

Kohavi, Ron, Longbotham, Roger, 2017. Online controlled experiments and A/B testing. Encycl. Mach. Learn. Data Min. 7 (8), 922–929.

Kohavi, Ron, Longbotham, Roger, Sommerfield, Dan, Henne, Randal, 2009. Controlled experiments on the web: Survey and practical guide. Data Min. Knowl. Discov. 18, 140–181. http://dx.doi.org/10.1007/s10618-008-0114-1.

Kohavi, Ron, Tang, Diane, Xu, Ya, 2020. Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing. Cambridge University Press, Cambridge, United Kingdom, http://dx.doi.org/10.1017/9781108653985.

Kohavi, Ron, Thomke, Stefan, 2017. The surprising power of online experiments. Harvard Business Review 74–82.

Kornilova, Anastasiia, Bernardi, Lucas, 2021. Mining the stars: Learning quality ratings with user-facing explanations for vacation rentals. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining. WSDM '21, Association for Computing Machinery, New York, NY, USA, pp. 976–983. http://dx.doi.org/10.1145/3437963.3441812.

Koukouvis, Kostantinos, Cubero, Roberto Alcañiz, Pelliccione, Patrizio, 2016. A/B testing in E-commerce sales processes. In: Crnkovic, Ivica, Troubitsyna, Elena (Eds.), Software Engineering for Resilient Systems. Springer International Publishing, Cham, pp. 133–148.

Kumar, Ratnakar, Hasteer, Nitasha, 2017. Evaluating usability of a web application: A comparative analysis of open-source tools. In: 2017 2nd International Conference on Communication and Electronics Systems. ICCES, pp. 350–354. http://dx.doi.org/10.1109/CESYS.2017.8321296.

Kumar, Anuj, Hosanagar, Kartik, 2017. Measuring the value of recommendation links on product demand. SSRN Electron. J. http://dx.doi.org/10.2139/ssrn.2909971.

Lalmas, Mounia, Lehmann, Janette, Shaked, Guy, Silvestri, Fabrizio, Tolomei, Gabriele, 2015. Promoting positive post-click experience for in-stream yahoo gemini users. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '15, Association for Computing Machinery, New York, NY, USA, pp. 1929–1938. http://dx.doi.org/10.1145/2783258.2788581.

Lee, Minyong R., Shen, Milan, 2018. Winner's curse: Bias estimation for total effects of features in online controlled experiments. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. KDD '18, Association for Computing Machinery, New York, NY, USA, pp. 491–499. http://dx.doi.org/10.1145/3219819.3219905.

Lettner, Florian, Holzmann, Clemens, Hutflesz, Patrick, 2013. Enabling A/B testing of native mobile applications by remote user interface exchange. In: Moreno-Díaz, Roberto, Pichler, Franz, Quesada-Arencibia, Alexis (Eds.), Computer Aided Systems Theory - EUROCAST 2013. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 458–466.

Li, Paul Luo, Chai, Xiaoyu, Campbell, Frederick, Liao, Jilong, Abburu, Neeraja, Kang, Minsuk, Niculescu, Irina, Brake, Greg, Patil, Siddharth, Dooley, James, Paddock, Brandon, 2021a. Evolving software to be ML-driven utilizing real-world A/B testing: Experiences, insights, challenges. In: 2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP). pp. 170–179. http://dx.doi.org/10.1109/ICSE-SEIP52600.2021.00026.

Li, Paul Luo, Dmitriev, Pavel, Hu, Huibin Mary, Chai, Xiaoyu, Dimov, Zoran, Paddock, Brandon, Li, Ying, Kirshenbaum, Alex, Niculescu, Irina, Thoresen, Taj, 2019. Experimentation in the operating system: The windows experimentation platform. In: 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP). pp. 21–30. http://dx.doi.org/10.1109/ICSE-SEIP.2019.00011.

Li, Lihong, Kim, Jin Young, Zitouni, Imed, 2015. Toward predicting the outcome of an A/B experiment for search relevance. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. WSDM '15, Association for Computing Machinery, New York, NY, USA, pp. 37–46. http://dx.doi.org/10.1145/2684822.2685311.

Li, Yiyang, Tao, Guanyu, Zhang, Weinan, Yu, Yong, Wang, Jun, 2017. Content recommendation by noise contrastive transfer learning of feature representation. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. CIKM '17, Association for Computing Machinery, New York, NY, USA, pp. 1657–1665. http://dx.doi.org/10.1145/3132847.3132855.

Li, Ye, Xie, Hong, Lin, Yishi, Lui, John C.S., 2021b. Unifying offline causal inference and online bandit learning for data driven decision. In: Proceedings of the Web Conference 2021. WWW '21, Association for Computing Machinery, New York, NY, USA, pp. 2291–2303. http://dx.doi.org/10.1145/3442381.3449982.

Li, Hannah, Zhao, Geng, Johari, Ramesh, Weintraub, Gabriel Y., 2022. Interference, bias, and variance in two-sided marketplace experimentation: Guidance for platforms. In: Proceedings of the ACM Web Conference 2022. WWW '22, Association for Computing Machinery, New York, NY, USA, pp. 182–192. http://dx.doi.org/10.1145/3485447.3512063.

Li, Chengbo, Zhu, Lin, Fu, Guangyuan, Du, Longzhi, Zhao, Canhua, Ma, Tianlun, Ye, Chang, Lee, Pei, 2021c. Learning to bundle proactively for on-demand meal delivery. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. CIKM '21, Association for Computing Machinery, New York, NY, USA, pp. 3898–3905. http://dx.doi.org/10.1145/3459637.3481931.

Liou, Kevin, Taylor, Sean J., 2020. Variance-weighted estimators to improve sensitivity in online experiments. In: Proceedings of the 21st ACM Conference on Economics and Computation. EC '20, Association for Computing Machinery, New York, NY, USA, pp. 837–850. http://dx.doi.org/10.1145/3391403.3399542.

Liu, Sophia, Fabijan, Aleksander, Furchtgott, Michael, Gupta, Somit, Janowski, Pawel, Qin, Wen, Dmitriev, Pavel, 2019. Enterprise-level controlled experiments at scale: Challenges and solutions. In: 2019 45th Euromicro Conference on Software Engineering and Advanced Applications. SEAA, pp. 29–37. http://dx.doi.org/10.1109/SEAA.2019.00013.

Liu, Yuchu, Mattos, David Issa, Bosch, Jan, Olsson, Helena Holmström, Lantz, Jonn, 2021. Size matters? Or not: A/b testing with limited sample in automotive embedded software. In: 2021 47th Euromicro Conference on Software Engineering and Advanced Applications. SEAA, pp. 300–307. http://dx.doi.org/10.1109/SEAA53835.2021.00046.

Machmouchi, Widad, Awadallah, Ahmed Hassan, Zitouni, Imed, Buscher, Georg, 2017. Beyond success rate: Utility as a search quality metric for online experiments. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. CIKM '17, Association for Computing Machinery, New York, NY, USA, pp. 757–765. http://dx.doi.org/10.1145/3132847.3132850.

Madeyski, Lech, Orzeszyna, Wojciech, Torkar, Richard, Józala, Mariusz, 2014. Overcoming the equivalent mutant problem: A systematic literature review and a comparative experiment of second order mutation. IEEE Trans. Softw. Eng. 40 (1), 23–42. http://dx.doi.org/10.1109/TSE.2013.44.

Madlberger, Maria, Jizdny, Jiri, 2021. Impact of promotional social media content on click-through rate - evidence from a FMCG company. In: 20th International Conferences on WWW/Internet 2021 and Applied Computing 2021. pp. 3–10, URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85124068035&partnerID=40&md5=c0b8f49a3b48b3d561fd0ed305eb1895, Cited by: 0.

Mahdavi-Hezavehi, Sara, Durelli, Vinicius H.S., Weyns, Danny, Avgeriou, Paris, 2017. A systematic literature review on methods that handle multiple quality attributes in architecture-based self-adaptive systems. Inf. Softw. Technol. 90, 1–26. http://dx.doi.org/10.1016/j.infsof.2017.03.013.

Masuda, Taisei, Murakami, Kyoko, Sugiura, Kenkichi, Sakui, Sho, Philip Schuring, Ron, Mori, Mitsuhiro, 2022. A phase 1/2 randomised placebo-controlled study of the COVID-19 vaccine mRNA-1273 in healthy Japanese adults: An interim report. Vaccine 40 (13), 2044–2052. http://dx.doi.org/10.1016/j.vaccine.2022.02.030.

Mattos, David Issa, Bosch, Jan, Holmström Olsson, Helena, 2017a. More for less: Automated experimentation in software-intensive systems. In: Felderer, Michael, Méndez Fernández, Daniel, Turhan, Burak, Kalinowski, Marcos, Sarro, Federica, Winkler, Dietmar (Eds.), Product-Focused Software Process Improvement. Springer International Publishing, Cham, pp. 146–161.

Mattos, David Issa, Bosch, Jan, Olsson, Helena Holmström, 2017b. Your system gets better every day you use it: Towards automated continuous experimentation. In: 2017 43rd Euromicro Conference on Software Engineering and Advanced Applications. SEAA, pp. 256–265. http://dx.doi.org/10.1109/SEAA.2017.15.

Mattos, David Issa, Bosch, Jan, Olsson, Helena Holmström, 2018. Challenges and strategies for undertaking continuous experimentation to embedded systems: Industry and research perspectives. In: Garbajosa, Juan, Wang, Xiaofeng, Aguiar, Ademar (Eds.), Agile Processes in Software Engineering and Extreme Programming. Springer International Publishing, Cham, pp. 277–292. http://dx.doi.org/10.1007/978-3-319-91602-6_20.

Mattos, David Issa, Bosch, Jan, Olsson, Helena Holmstrom, Maryam Korshani, Aita, Lantz, Jonn, 2020. Automotive A/B testing: Challenges and lessons learned from practice. In: 2020 46th Euromicro Conference on Software Engineering and Advanced Applications. SEAA, pp. 101–109. http://dx.doi.org/10.1109/SEAA51224.2020.00026.

Metrikov, Pavel, Diaz, Fernando, Lahaie, Sebastien, Rao, Justin, 2014. Whole page optimization: How page elements interact with the position auction. In: Proceedings of the Fifteenth ACM Conference on Economics and Computation. EC '14, Association for Computing Machinery, New York, NY, USA, pp. 583–600. http://dx.doi.org/10.1145/2600057.2602871.

Miikulainen, Risto, Brundage, Myles, Epstein, Jonathan, Foster, Tyler, Hodjat, Babak, Iscoe, Neil, Jiang, Jingbo, Legrand, Diego, Nazari, Sam, Qiu, Xin, Scharff, Michael, Schoolland, Cory, Severn, Robert, Shagrin, Aaron, 2020. Ascend by evolv: AI-based massively multivariate conversion rate optimization. AI Mag. 41 (1), 44–60. http://dx.doi.org/10.1609/aimag.v41i1.5256.

Okoshi, Tadashi, Tsubouchi, Kota, Tokuda, Hideyuki, 2019. Real-world product deployment of adaptive push notification scheduling on smartphones. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. KDD '19, Association for Computing Machinery, New York, NY, USA, pp. 2792–2800. http://dx.doi.org/10.1145/3292500.3330732.

Ozawa, Takumi, Sekiguchi, Akiyuki, Tsuda, Kazuhiko, 2016. A method for the construction of user targeting knowledge for B2B industry website. Procedia Comput. Sci. 96, 1147–1155. http://dx.doi.org/10.1016/j.procs.2016.08.157, Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 20th International Conference KES-2016.

Pelleg, Dan, Rokhlenko, Oleg, Szpektor, Idan, Agichtein, Eugene, Guy, Ido, 2016. When the crowd is not enough: Improving user experience with social media through automatic quality analysis. In: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing. CSCW '16, Association for Computing Machinery, New York, NY, USA, pp. 1080–1090. http://dx.doi.org/10.1145/2818048.2820022.

Peska, Ladislav, Vojtas, Peter, 2020. Off-line vs. On-line evaluation of recommender systems in small E-commerce. In: Proceedings of the 31st ACM Conference on Hypertext and Social Media. HT '20, Association for Computing Machinery, New York, NY, USA, pp. 291–300. http://dx.doi.org/10.1145/3372923.3404781.

Poyarkov, Alexey, Drutsa, Alexey, Khalyavin, Andrey, Gusev, Gleb, Serdyukov, Pavel, 2016. Boosted decision tree regression adjustment for variance reduction in online controlled experiments. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16, Association for Computing Machinery, New York, NY, USA, pp. 235–244. http://dx.doi.org/10.1145/2939672.2939688.

Qu, Jia, Zhang, Jing, 2016. Validating mobile designs with agile testing in China: Based on baidu map for mobile. In: Marcus, Aaron (Ed.), Design, User Experience, and Usability: Design Thinking and Methods. Springer International Publishing, Cham, pp. 491–498.

Quin, Federico, Weyns, Danny, Galster, Matthias, 2023. Study systematic literature review on A/B testing. URL https://people.cs.kuleuven.be/danny.weyns/material/SLR_AB/.

Renz, Jan, Hoffmann, Daniel, Staubitz, Thomas, Meinel, Christoph, 2016. Using A/B testing in MOOC environments. In: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge. LAK '16, Association for Computing Machinery, New York, NY, USA, pp. 304–313. http://dx.doi.org/10.1145/2883851.2883876.

Reza, Mohi, Kim, Juho, Bhattacharjee, Ananya, Rafferty, Anna N., Williams, Joseph Jay, 2021. The mooclet framework: Unifying experimentation, dynamic improvement, and personalization in online courses. In: Proceedings of the Eighth ACM Conference on Learning @ Scale. In: L@S '21, Association for Computing Machinery, New York, NY, USA, pp. 15–26. http://dx.doi.org/10.1145/3430895.3460128.

Rodríguez, Pilar, Haghighatkhah, Alireza, Lwakatare, Lucy Ellen, Teppola, Susanna, Suomalainen, Tanja, Eskeli, Juho, Karvonen, Teemu, Kuvaja, Pasi, Verner, June M., Oivo, Markku, 2017. Continuous deployment of software intensive products and services: A systematic mapping study. J. Syst. Softw. 123, 263–291. http://dx.doi.org/10.1016/j.jss.2015.12.015.

Ros, Rasmus, Runeson, Per, 2018. Continuous experimentation and A/B testing: A mapping study. In: Proceedings of the 4th International Workshop on Rapid Continuous Software Engineering. RCoSE '18, Association for Computing Machinery, New York, NY, USA, pp. 35–41. http://dx.doi.org/10.1145/3194760.3194766.

Rosenfeld, Nir, Mansour, Yishay, Yom-Tov, Elad, 2017. Predicting counterfactuals from large historical data and small randomized trials. In: Proceedings of the 26th International Conference on World Wide Web Companion. In: WWW '17 Companion, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, pp. 602–609. http://dx.doi.org/10.1145/3041021.3054190.

Sajeev, Sandra, Huang, Jade, Karampatziakis, Nikos, Hall, Matthew, Kochman, Sebastian, Chen, Weizhu, 2021. Contextual bandit applications in a customer support bot. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. KDD '21, Association for Computing Machinery, New York, NY, USA, pp. 3522–3530. http://dx.doi.org/10.1145/3447548.3467165.

Satyal, Suhrid, Weber, Ingo, Paik, Hye-young, Di Ciccio, Claudio, Mendling, Jan, 2017. AB-BPM: Performance-driven instance routing for business process improvement. In: Carmona, Josep, Engels, Gregor, Kumar, Akhil (Eds.), Business Process Management. Springer International Publishing, Cham, pp. 113–129.

Satyal, Suhrid, Weber, Ingo, young Paik, Hye, Di Ciccio, Claudio, Mendling, Jan, 2019. Business process improvement with the AB-BPM methodology. Inf. Syst. 84, 283–298. http://dx.doi.org/10.1016/j.is.2018.06.007.

Saveski, Martin, Pouget-Abadie, Jean, Saint-Jacques, Guillaume, Duan, Weitao, Ghosh, Souvik, Xu, Ya, Airoldi, Edoardo M., 2017. Detecting network effects: Randomizing over randomized experiments. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '17, Association for Computing Machinery, New York, NY, USA, pp. 1027–1035. http://dx.doi.org/10.1145/3097983.3098192.

Schermann, Gerald, Schöni, Dominik, Leitner, Philipp, Gall, Harald C., 2016. Bifrost: Supporting continuous deployment with automated enactment of multi-phase live testing strategies. In: Proceedings of the 17th International Middleware Conference. Middleware '16, Association for Computing Machinery, New York, NY, USA, http://dx.doi.org/10.1145/2988336.2988348.

Shariat, Shahriar, Orten, Burkay, Dasdan, Ali, 2017. Online evaluation of bid prediction models in a large-scale computational advertising platform: Decision making and insights. Knowl. Inf. Syst. 51 (1), 37–60. http://dx.doi.org/10.1007/s10115-016-0972-6.

Shi, Fanjuan, Ghedira, Chirine, Marini, Jean-Luc, 2015. Context adaptation for smart recommender systems. IT Prof. 17 (6), 18–26. http://dx.doi.org/10.1109/MITP.2015.96.

Siegmund, Janet, Siegmund, Norbert, Apel, Sven, 2015. Views on internal and external validity in empirical software engineering. In: 2015 IEEE/ACM 37th IEEE International Conference on Software Engineering. Vol. 1, IEEE Press, Los Alamitos, CA, USA, pp. 9–19. http://dx.doi.org/10.1109/ICSE.2015.24.

Silberstein, Natalia, Somekh, Oren, Koren, Yair, Aharon, Michal, Porat, Dror, Shahar, Avi, Wu, Tingyi, 2020. Ad close mitigation for improved user experience in native advertisements. In: Proceedings of the 13th International Conference on Web Search and Data Mining. WSDM '20, Association for Computing Machinery, New York, NY, USA, pp. 546–554. http://dx.doi.org/10.1145/3336191.3371798.

Siqueira, Jorge Gabriel, de Paula, Melise M.V., 2018. IPEAD A/B test execution framework. In: Proceedings of the XIV Brazilian Symposium on Information Systems. SBSI '18, Association for Computing Machinery, New York, NY, USA, http://dx.doi.org/10.1145/3229345.3229360.

Siroker, Dan, Koomen, Pete, 2013. A/B Testing: The Most Powerful Way to Turn Clicks Into Customers, first ed. Wiley Publishing, Hoboken, NJ, USA.

Spang, Bruce, Hannan, Veronica, Kunamalla, Shravya, Huang, Te-Yuan, McKeown, Nick, Johari, Ramesh, 2021. Unbiased experiments in congested networks. In: Proceedings of the 21st ACM Internet Measurement Conference. IMC '21, Association for Computing Machinery, New York, NY, USA, pp. 80–95. http://dx.doi.org/10.1145/3487552.3487851.

Sriraman, Akshitha, Dhanotia, Abhishek, Wenisch, Thomas F., 2019. SoftSKU: Optimizing server architectures for microservice diversity @scale. In: 2019 ACM/IEEE 46th Annual International Symposium on Computer Architecture. ISCA, pp. 513–526.

Stewart, Michael, Hodkiewicz, Melinda, Li, Sirui, 2023. Large language models for failure mode classification: An investigation. arXiv:2309.08181.

Sun, Fei, Jiang, Peng, Sun, Hanxiao, Pei, Changhua, Ou, Wenwu, Wang, Xiaobo, 2018. Multi-source pointer network for product title summarization. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management. CIKM '18, Association for Computing Machinery, New York, NY, USA, pp. 7–16. http://dx.doi.org/10.1145/3269206.3271722.

Szpektor, Idan, Maarek, Yoelle, Pelleg, Dan, 2013. When relevance is not enough: Promoting diversity and freshness in personalized question recommendation. In: Proceedings of the 22nd International Conference on World Wide Web. WWW '13, Association for Computing Machinery, New York, NY, USA, pp. 1249–1260. http://dx.doi.org/10.1145/2488388.2488497.

Tagami, Yukihiro, Hotta, Toru, Tanaka, Yusuke, Ono, Shingo, Tsukamoto, Koji, Tajima, Akira, 2014. Filling context-ad vocabulary gaps with click logs. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '14, Association for Computing Machinery, New York, NY, USA, pp. 1955–1964. http://dx.doi.org/10.1145/2623330.2623334.

Tamburrelli, Giordano, Margara, Alessandro, 2014. Towards automated A/B testing. In: Le Goues, Claire, Yoo, Shin (Eds.), Search-Based Software Engineering. Springer International Publishing, Cham, pp. 184–198.

Tang, Diane, Agarwal, Ashish, O'Brien, Deirdre, Meyer, Mike, 2010. Overlapping experiment infrastructure: More, better, faster experimentation. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '10, Association for Computing Machinery, New York, NY, USA, pp. 17–26. http://dx.doi.org/10.1145/1835804.1835810.

Toslali, Mert, Parthasarathy, Srinivasan, Oliveira, Fabio, Coskun, Ayse K., 2020. JACKPOT: Online experimentation of cloud microservices. In: HotCloud 2020 - 12th USENIX Workshop on Hot Topics in Cloud Computing, co-located with USENIX ATC 2020. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85091892156&partnerID=40&md5=cae12fe24f34f2bb0818e448f8c07fbf, Cited by: 1.

Tsigkanos, Christos, Rani, Pooja, Müller, Sebastian, Kehrer, Timo, 2023. Variable discovery with large language models for metamorphic testing of scientific software. In: Computational Science – ICCS 2023. Springer Nature Switzerland, Cham, pp. 321–335.

Tu, Ye, Basu, Kinjal, DiCiccio, Cyrus, Bansal, Romil, Nandy, Preetam, Jaikumar, Padmini, Chatterjee, Shaunak, 2021. Personalized treatment selection using causal heterogeneity. In: Proceedings of the Web Conference 2021. WWW '21, Association for Computing Machinery, New York, NY, USA, pp. 1574–1585. http://dx.doi.org/10.1145/3442381.3450075.

Ueoka, Yutaro, Tsubouchi, Kota, Shimizu, Nobuyuki, 2020. Tackling cannibalization problems for online advertisement. In: Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization. UMAP '20, Association for Computing Machinery, New York, NY, USA, pp. 358–362. http://dx.doi.org/10.1145/3340631.3394875.

Vanderdonckt, Jean, Zen, Mathieu, Vatavu, Radu-Daniel, 2019. AB4web: An on-line A/B tester for comparing user interface design alternatives. Proc. ACM Hum.-Comput. Interact. 3 (EICS), http://dx.doi.org/10.1145/3331160.

Vasthimal, Deepak Kumar, Srirama, Pavan Kumar, Akkinapalli, Arun Kumar, 2019. Scalable data reporting platform for A/B tests. In: 2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security. IDS, pp. 230–238. http://dx.doi.org/10.1109/BigDataSecurity-HPSC-IDS.2019.00052.

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Lukasz, Polosukhin, Illia, 2023. Attention is all you need. arXiv:1706.03762.

Walper, Daniel, Kassau, Julia, Methfessel, Philipp, Pronold, Timo, Einhauser, Wolfgang, 2020. Optimizing user interfaces in food production: gaze tracking is more sensitive for A-B-testing than behavioral data alone. In: ACM Symposium on Eye Tracking Research and Applications. In: ETRA '20 Short Papers, Association for Computing Machinery, New York, NY, USA, pp. 1–4. http://dx.doi.org/10.1145/3379156.3391351.

Wang, Zenan, Carrion, Carlos, Lin, Xiliang, Ji, Fuhua, Bao, Yongjun, Yan, Weipeng, 2022. Adaptive experimentation with delayed binary feedback. In: Proceedings of the ACM Web Conference 2022. WWW '22, Association for Computing Machinery, New York, NY, USA, pp. 2247–2255. http://dx.doi.org/10.1145/3485447.3512097.

Wang, Yu, Gupta, Somit, Lu, Jiannan, Mahmoudzadeh, Ali, Liu, Sophia, 2019. On heavy-user bias in A/B testing. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. CIKM '19, Association for Computing Machinery, New York, NY, USA, pp. 2425–2428. http://dx.doi.org/10.1145/3357384.3358143.

Wang, Jian, Hardtke, David, 2015. User latent preference model for better downside management in recommender systems. In: Proceedings of the 24th International Conference on World Wide Web. WWW '15, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, pp. 1209–1219. http://dx.doi.org/10.1145/2736277.2741126.

Wang, Weinan, Zhang, Xi, 2021. CONQ: Continuous quantile treatment effects for large-scale online controlled experiments. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining. WSDM '21, Association for Computing Machinery, New York, NY, USA, pp. 202–210. http://dx.doi.org/10.1145/3437963.3441779.

Wei, Jason, Tay, Yi, Bommasani, Rishi, Raffel, Colin, Zoph, Barret, Borgeaud, Sebastian, Yogatama, Dani, Bosma, Maarten, Zhou, Denny, Metzler, Donald, Chi, Ed H., Hashimoto, Tatsunori, Vinyals, Oriol, Liang, Percy, Dean, Jeff, Fedus, William, 2022. Emergent abilities of large language models. arXiv:2206.07682.

Wu, Liang, Grbovic, Mihajlo, 2020. How airbnb tells you will enjoy sunset sailing in Barcelona? recommendation in a two-sided travel marketplace. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '20, Association for Computing Machinery, New York, NY, USA, pp. 2387–2396. http://dx.doi.org/10.1145/3397271.3401444.

Wu, Jie J.W., Mazzuchi, Thomas A., Sarkani, Shahram, 2023a. Comparison of multi-criteria decision-making methods for online controlled experiments in a launch decision-making framework. Inf. Softw. Technol. 155, 107115. http://dx.doi.org/10.1016/j.infsof.2022.107115.

Wu, Jie J.W., Mazzuchi, Thomas A., Sarkani, Shahram, 2023b. A multi-objective evolutionary approach towards automated online controlled experiments. J. Syst. Softw. 203, 111703. http://dx.doi.org/10.1016/j.jss.2023.111703.

Wu, Yuhang, Zheng, Zeyu, Zhang, Guangyu, Zhang, Zuohua, Wang, Chu, 2022. Non-stationary A/B tests. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. KDD '22, Association for Computing Machinery, New York, NY, USA, pp. 2079–2089. http://dx.doi.org/10.1145/3534678.3539325.

Xia, Tong, Bhardwaj, Sumit, Dmitriev, Pavel, Fabijan, Aleksander, 2019. Safe velocity: A practical guide to software deployment at scale using controlled rollout. In: 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP). pp. 11–20. http://dx.doi.org/10.1109/ICSE-SEIP.2019.00010.

Xie, Yuxiang, Chen, Nanyu, Shi, Xiaolin, 2018. False discovery rate controlled heterogeneous treatment effect detection for online controlled experiments. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. KDD '18, Association for Computing Machinery, New York, NY, USA, pp. 876–885. http://dx.doi.org/10.1145/3219819.3219860.

Xie, Yuxiang, Xu, Meng, Chow, Evan, Shi, Xiaolin, 2021. How to measure your app: A couple of pitfalls and remedies in measuring app performance in online controlled experiments. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining. WSDM '21, Association for Computing Machinery, New York, NY, USA, pp. 949–957. http://dx.doi.org/10.1145/3437963.3441742.

Xu, Ya, Chen, Nanyu, 2016. Evaluating mobile apps with A/B and quasi A/B tests. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16, Association for Computing Machinery, New York, NY, USA, pp. 313–322. http://dx.doi.org/10.1145/2939672.2939703.

Xu, Ya, Chen, Nanyu, Fernandez, Addrian, Sinno, Omar, Bhasin, Anmol, 2015. From infrastructure to culture: A/B testing challenges in large scale social networks. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '15, Association for Computing Machinery, New York, NY, USA, pp. 2227–2236. http://dx.doi.org/10.1145/2783258.2788602.

Xu, Ye, Li, Zang, Gupta, Abhishek, Bugdayci, Ahmet, Bhasin, Anmol, 2014. Modeling professional similarity by mining professional career trajectories. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '14, Association for Computing Machinery, New York, NY, USA, pp. 1945–1954. http://dx.doi.org/10.1145/2623330.2623368.

Xu, Yanbo, Mahajan, Divyat, Manrao, Liz, Sharma, Amit, Kıcıman, Emre, 2021. Split-treatment analysis to rank heterogeneous causal effects for prospective interventions. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining. WSDM '21, Association for Computing Machinery, New York, NY, USA, pp. 409–417. http://dx.doi.org/10.1145/3437963.3441821.

Yaman, Sezin Gizem, Munezero, Myriam, Münch, Jürgen, Fagerholm, Fabian, Syd, Ossi, Aaltola, Mika, Palmu, Christina, Männistö, Tomi, 2017. Introducing continuous experimentation in large software-intensive product and service organisations. J. Syst. Softw. 133, 195–211. http://dx.doi.org/10.1016/j.jss.2017.07.009.

Yang, Wanshan, Yang, Gemeng, Huang, Ting, Chen, Lijun, Liu, Youjian Eugene, 2018. Whales, dolphins, or minnows? Towards the player clustering in free online games based on purchasing behavior via data mining technique. In: 2018 IEEE International Conference on Big Data (Big Data). pp. 4101–4108. http://dx.doi.org/10.1109/BigData.2018.8622067.

Ye, Runlong, Chen, Pan, Mao, Yini, Wang-Lin, Angela, Shaikh, Hammad, Zavaleta Bernuy, Angela, Williams, Joseph Jay, 2022. Behavioral consequences of reminder emails on students' academic performance: A real-world deployment. In: Proceedings of the 23rd Annual Conference on Information Technology Education. SIGITE '22, Association for Computing Machinery, New York, NY, USA, pp. 16–22. http://dx.doi.org/10.1145/3537674.3554740.

Yoneda, Takeshi, Kozawa, Shunsuke, Osone, Keisuke, Koide, Yukinori, Abe, Yosuke, Seki, Yoshifumi, 2019. Algorithms and system architecture for immediate personalized news recommendations. In: IEEE/WIC/ACM International Conference on Web Intelligence. WI '19, Association for Computing Machinery, New York, NY, USA, pp. 124–131. http://dx.doi.org/10.1145/3350546.3352509.

Young, Scott W.H., 2014. Improving library user experience with A/B testing: Principles and process. Weave: J. Library User Exp. 1, http://dx.doi.org/10.3998/weave.12535642.0001.101.

Yu, Miao, Lu, Wenbin, Song, Rui, 2020. A new framework for online testing of heterogeneous treatment effect. In: AAAI 2020 - 34th AAAI Conference on Artificial Intelligence. pp. 10310–10317, URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85106588123&partnerID=40&md5=53544f162212be7cd129e1f196debcd8, Cited by: 2.

Zhang, He, Ali Babar, Muhammad, 2010. On searching relevant studies in software engineering. In: Proceedings of the 14th International Conference on Evaluation and Assessment in Software Engineering. EASE '10, BCS Learning & Development Ltd., Swindon, GBR, pp. 111–120.

Zhao, Zhenyu, He, Yan, Chen, Miao, 2017. Inform product change through experimentation with data-driven behavioral segmentation. In: 2017 IEEE International Conference on Data Science and Advanced Analytics. DSAA, pp. 69–78. http://dx.doi.org/10.1109/DSAA.2017.65.

Zhu, Xizhou, Chen, Yuntao, Tian, Hao, Tao, Chenxin, Su, Weijie, Yang, Chenyu, Huang, Gao, Li, Bin, Lu, Lewei, Wang, Xiaogang, Qiao, Yu, Zhang, Zhaoxiang, Dai, Jifeng, 2023. Ghost in the minecraft: Generally capable agents for open-world environments via large language models with text-based knowledge and memory. arXiv:2305.17144.

Federico Quin obtained his Ph.D. from KU Leuven for research on machine learning in self-adaptive systems. He is currently working as a software engineer in the software industry. His research interests lie in software engineering aspects of A/B testing and self-adaptive systems.

Danny Weyns is a professor at Linnaeus University, Sweden, and the Katholieke Universiteit Leuven, Belgium. His current research interests lie in software engineering of self-adaptive systems.

Matthias Galster is a professor at the University of Canterbury in Christchurch, New Zealand.

Camila Costa Silva obtained her Ph.D. in Computer Science at the University of Canterbury, New Zealand and her M.Sc. in Information Systems at the University of Sao Paulo, Brazil. She is currently working as a research assistant.