

A/B Testing: A Systematic Literature Review

Federico Quin^{a,*}, Danny Weyns^{a,b}, Matthias Galster^c, Camila Costa Silva^c

^a*Distrinet, KU Leuven, Celestijnenlaan 200A, Leuven, 3000, Belgium*

^b*Linnaeus University, Universitetsplatsen 1, Växjö, 351 06, Sweden*

^c*University of Canterbury, 69 Creyke Road, Christchurch, 8140, New Zealand*

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 141 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.

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

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 [92, 58, 106, 71]. 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 [82, 161, 48]. A/B testing is extensively

*Corresponding author

Email addresses: federico.quin@kuleuven.be (Federico Quin), danny.weyns@kuleuven.be (Danny Weyns), matthias.galster@canterbury.ac.nz (Matthias Galster), camila.costasilva@pg.canterbury.ac.nz (Camila Costa Silva)

used in practice, including large and popular tech companies such as Google, Meta, LinkedIn, and Microsoft [77, 161, 104, 168].

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 [13, 12, 133] 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 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 [84]. 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 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 [145]. 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. Figure 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 [47].

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 [75, 156]. 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 [32]. 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

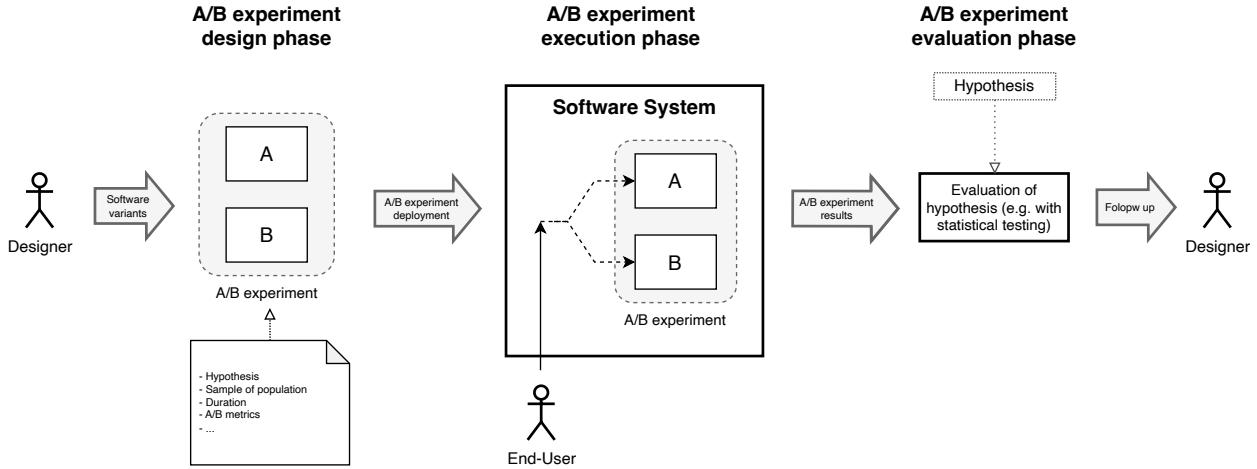


Figure 1: General A/B testing process.

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 [31], pharmaceutics [116], education [32], and nowadays also in software engineering [142, 34, 68].

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 [94, 59]. 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) [130], improving user interface (UI) design [158], improving user click rates [4], or evaluating non functional requirements in distributed services [14].

2.1.2. DevOps and A/B testing

Development Operations (DevOps in short) has gained popularity in recent years [35]. 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) [80]. 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 [172] 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. [58] 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 study 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. Rodriguez et al. [132] 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 [133] 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 [12] 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. [13] presented a systematic literature review on A/B testing and continuous experimentation, leveraging the results from previous mapping studies [133, 12]. 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. [53] 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. [61] 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. [62] 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. [119] 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. [69] 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 [17]:

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 [84]. 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 [129], which 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 studies¹, 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 [129].

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?

¹We 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.

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 [87, 73, 93, 92, 85, 47]. We then identified and applied a gold standard [178] to tune the terms. For a detailed description of the relevant terms and application of the gold standard, we refer to the research protocol [129]. Figure 2 (top) displays the final search query after applying the gold standard.

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 resulted in 3,944 papers, as shown in Figure 2. After filtering duplicate papers and selecting only journal versions of extensions of conference versions, 2,379 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.

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 (≤ 4 pages)², demos, extended abstracts, keynote talks, and tutorials;

²Papers published in the *Lecture Notes in Computer Science* format with < 8 pages are also considered short.

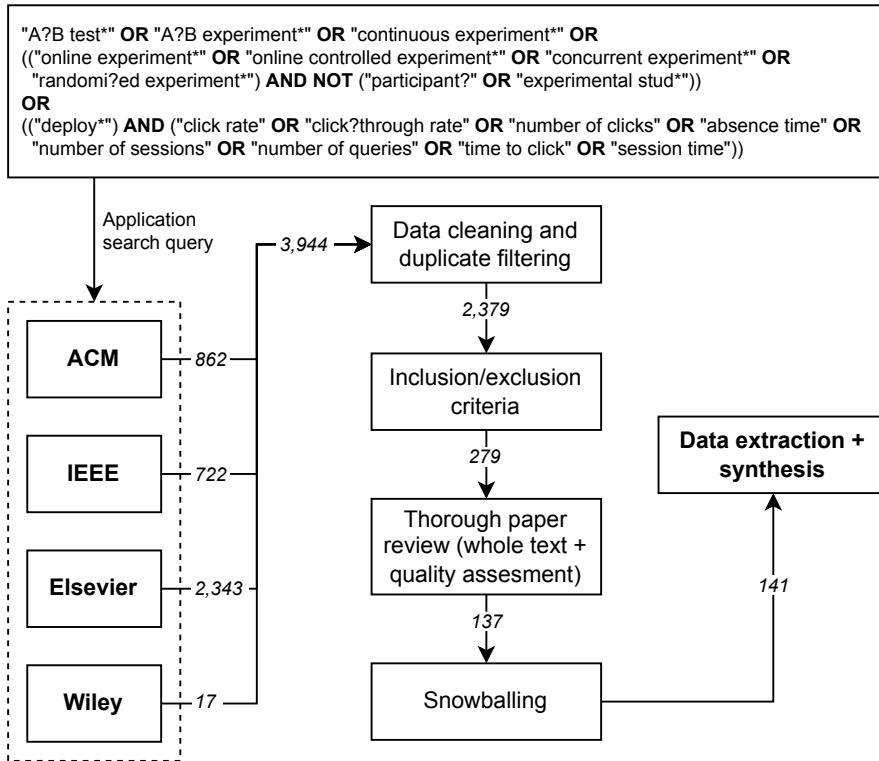


Figure 2: Primary studies selected for the systematic literature review.

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 137. 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 141, as shown in Figure 2.

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.

D1-4: Authors, year, title, and venue used for documentation purposes.

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	Paper type	Documentation
D6	Authors sector	Documentation
D7	Quality score	Documentation
D8	Application domain	RQ1
D9	A/B target	RQ1
D10	A/B test type	RQ2
D11	Used metrics	RQ2
D12	Statistical methods employed	RQ2
D13	Role of stakeholders in the experiment design	RQ2
D14	Additional data collected	RQ3
D15	Evaluation method	RQ3
D16	Use of test results	RQ3
D17	Role of stakeholder in experiment execution	RQ3
D18	Open problems	RQ4

- D5:** 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).
- D6:** 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.³
- D7:** A quality score for the reporting of the research [115]. 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 [113].
- D8:** 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 derived during data collection.
- D9:** 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.
- D10:** 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 A/B tests, Automated sequence of classic A/B tests, Automated sequence of multivariate A/B tests. Additional options were derived during data collection.
- D11:** 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.

³Academic refers to affiliations that are eligible to graduate master and/or PhD students.

Table 2: Paper types of the primary studies.

Type	Number of occurrences
Focus	90
Applied	51

- D12:** 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.
- D13:** 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.
- D14:** 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.
- D15:** The evaluation method used in the primary study⁴. Initial options include Illustrative example, Simulation, Empirical evaluation.
- D16:** 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.
- D17:** 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.
- D18:** 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 Paper type (**D5**), Authors sector (**D6**), and Quality score (**D7**).

Of the 141 primary studies, 90 (63.8%) have a focus on A/B testing itself, while 51 (36.2%) apply A/B testing or use it for evaluation purpose, see Table 2.

A majority of 72 primary studies (51.1%) have industrial authors, see Table 3. Forty-three studies (30.5%) have a mix of industry and academic authors, and 26 studies (18.4%) are from academic authors only.

Figure 3 shows the distribution of quality scores with an average of 8.81 [± 1.58]. 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.

⁴We 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 3: Author backgrounds of the primary studies.

Background	Number of occurrences
Academic	26
Industry	72
Mixed	43

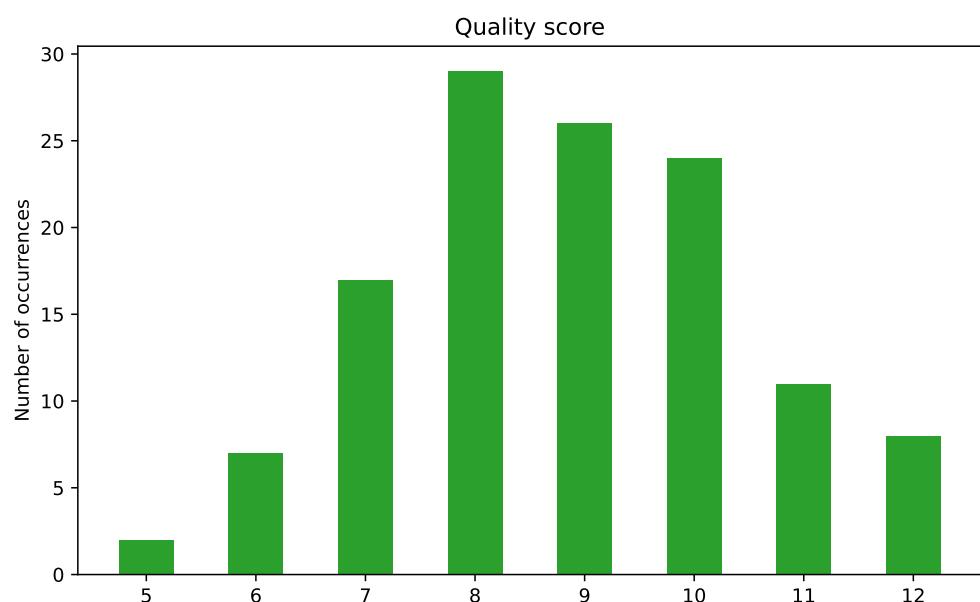


Figure 3: Quality scores of the primary studies.

Table 4: Identified application domains for A/B testing.

Application domain	Number of occurrences
Web	38
Search engine	35
E-commerce	27
Interaction	22
Finances	16
Transportation	4
Other	8
N/A	9

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 (**D8**), and A/B target (**D9**).

Application domain. Table 4 lists the application domains of the primary studies. The average number of domains is 1.13 (131 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 [109] or LinkedIn [170], news publishers [175, 60], and multimedia services, such as movie streaming at Netflix [9]. The second most popular domain is search engines (35 occurrences), with studies conducted at Yandex [46, 45], Bing [41, 112], Yahoo [6, 150], among others. A/B testing is also actively applied in E-commerce (27 occurrences), with examples from retail giant Amazon [52], the fashion industry [26], and C2C (consumer-to-consumer) businesses, such as Etsy [83] and Facebook marketplace [77]. Next we observe the application of A/B testing in what we group under "interaction" (22 occurrences), with digital communication software, such as Snap [167] and Skype [60], user-operating system interaction [74, 56], and application software, such as an App store [33] and mobile games [173]. Lastly, we note the financial application domain (16 occurrences), including studies at Yahoo finance [179] and Alipay [24], transportation (4 occurrences) at for instance Didi Chuxing [66]. Other domains are education (3 occurrences) [131] and robotics (2 occurrences) [118], 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 1.21 (120 primary studies applied A/B testing to one element, 26 studies to two elements, and 24 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 86.2% of all A/B targets reported in the primary studies. Notable, 32 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 [166].

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

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 [175], a search ranking algorithm [86], or an ad serving algorithm [16].	58
Visual elements	Change to visual components such as updates to a website layout [22] or a general user interface update [40].	33
Workflow / process	Alteration to the workflow of an application, e.g. the addition of a feedback button to a dashboard [110], or a change in a user workflow, e.g. the process of a virtual assistant tool [96].	28
Back-end	Optimization of a software component that is not directly visible to the user, such as testing server optimizations [127] or adjusting application parameters for better performance [60].	10
New application functionality	Newly introduced functionality, such as a new widget on a web-page [28] or additional content that is presented to the user after performing a search query [112].	6
Other	This category comprises three other A/B targets: different timing and content of emails sent [174], varying educational resources presented to the user [131], and the page configuration of a website [157].	3
Unspecified	The target of the A/B test was not specified in the study.	32

Table 6: Application domain \times A/B target

<i>Application domain</i>	<i>A/B target</i>	Algorithm	Visual elements	Workflow / process	Back-end	New app. func.	Other
Web	17	6	8	1	3	0	
Search engine	17	16	3	6	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	

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 [90] and changing the position of advertisements on the result page [121].
- 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 [162], and changes to the order assignment policy for on-demand meal delivery platforms [102].
- 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 [147]. 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 [18], one study evaluated a change in advertisement strategies [75], the last study evaluated the option to pay for "sponsored search" (to prioritize search results) [19].
- 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 [103].
- 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.

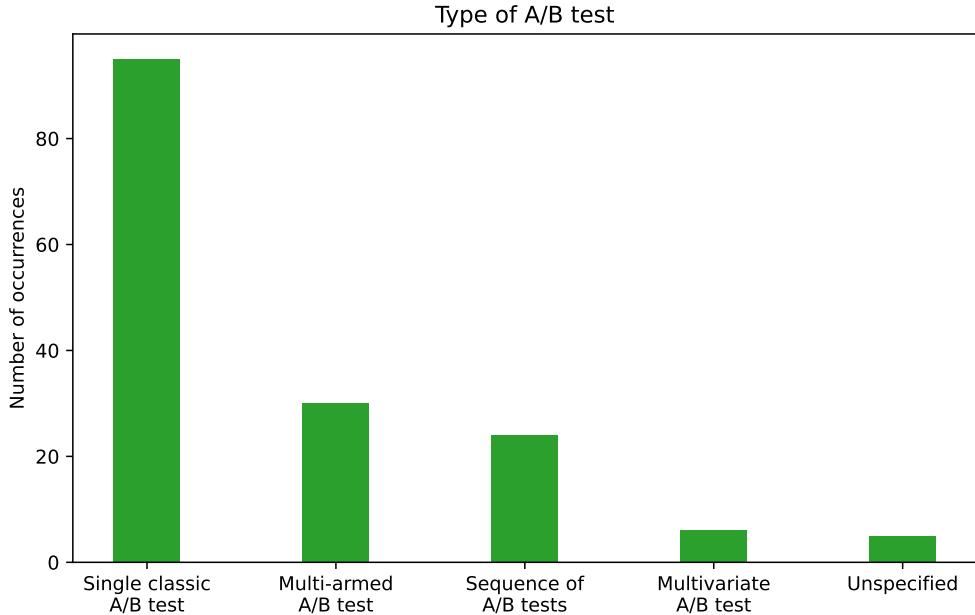


Figure 4: Identified A/B test types.

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 (**D10**), Used metrics (**D11**), Statistical methods employed (**D12**), and Role of stakeholders in the experiment design (**D13**).

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.

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. Figure 4 shows the frequencies of these different A/B test types extracted from the primary studies.

Overall, we identified 155 occurrences of A/B test types, i.e., an average of 1.13 occurrences per primary study (123 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 (95 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., [141, 149], and to test different advertisement serving algorithms, e.g., [155]. 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) [63], manually testing multiple iterations of machine learning algorithms sequentially [105], and automatically executing a sequence of A/B tests to handle controlled feature release in DevOps [139]. The last identified A/B test type is multivariate A/B test (6 occurrences). This type of test evaluates various combinations of multiple

Table 7: Identified A/B metrics.

A/B metric	Number of occurrences
Engagement metrics	225
Click metrics	82
Monetary metrics	64
Performance metrics	50
Negative metrics	34
View metrics	21
Feedback metrics	17

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 [40].

A/B metrics.. Table 7 lists the A/B metrics that we extracted from the primary studies. In total, 493 occurrences of A/B metrics were reported in the primary studies. With a total of 198 experiments spread over 141 studies, this gives an average of 2.12 metrics per experiment⁵ (ranging from 1 to 8 metrics per experiment). The most common group of A/B metrics are engagement metrics (225 occurrences) that refer to the number of conversions⁶, 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).⁷ The second largest group are click metrics (82 occurrences). Examples include number of clicks, clicks per query, and good click rate⁸. 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 (94 occurrences in total). The main test used in this group is a student t-test, e.g. [75, 71]. Other tests in this group are the Kolmogorov-Smirnov test, e.g., [140], Mann-Whitney test, e.g., [137], and Wilcoxon signed-rank test, e.g., [156]. 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 results⁹. The second most commonly used method is bootstrapping (11 occurrences). This method constructs multiple datasets by resampling the original dataset [46]. 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 [110]. The third mostly commonly used statistical method is a hypothesis test that tests for inference and goodness of fit

⁵We excluded experiments and corresponding metrics of primary studies that analyzed a large number of previously conducted A/B tests.

⁶A conversion is a desired action taken in the A/B test.

⁷Note 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.

⁸Good clicks are described as clicks that are meaningful during the search query session [20].

⁹However, 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 8: Statistical methods employed during A/B testing.

Statistical methods employed	Number of occurrences
Hypothesis - equality	57
Hypothesis - equality (concrete method unspecified)	37
Bootstrapping	11
Hypothesis - inference	8
Goodness of fit	8
Correction method	7
Estimator	6
Hypothesis - independence	5
Regression method	2

(both 8 occurrences). Examples of inference hypothesis tests include using Bayesian analysis approach to ensure multiple simultaneously running experiments do not interfere [89], and a Bayesian approach to infer the causal effect of running ad campaigns [15]. Examples of goodness of fit methods include sequential testing methods that are based on likelihood ratio tests [83], and a Wald test [81]. The remaining groups are correction methods (7 occurrences) with e.g. Bonferroni correction [177]; custom estimators for observations in A/B testing (6 occurrences), e.g., an estimator that takes variances into account [109]; hypothesis tests for independence (5 occurrences), containing χ^2 tests [150]; and regression methods (2 occurrences), e.g. CUPED [48].

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 (127 occurrences), experiment architect (111 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 (60 occurrences). Finally, the main task of the setup technician is performing post-design activities of A/B tests (25 occurrences).

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).

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 (127)	Design and tune variants	Designing and tuning the variants to test. Examples are tweaking the A/B variants [141], or designing A/B variants for different kind of populations (e.g., old vs new users) [21].	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 [50] or the specification of a pre-determined hypothesis for the A/B test [5].	48
	Perform pre-design actions	Actions that are taken before designing the A/B test. Examples include providing motivation for A/B tests [157] or performing offline A/B tests before moving to online A/B testing [72].	12
Experiment architect (111)	Determine duration	Determining the duration of the A/B test. Examples include choosing a fixed experiment duration (e.g., 1 week) [5] or via an explicit expiration date [106].	60
	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 [163], an assignment where the target population is determined over a two week period [173], or an assignment where network effects have to be taken into account [102].	51
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 [179, 33], validation of the A/B test design [110], or scheduling the execution of the A/B test [157].	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 [112].	6

Table 10: Tasks of stakeholders \times A/B test type

<i>Test type (total occ.)</i>	Single classic A/B test (95)	Multi-armed A/B test (30)	Sequence of A/B tests (24)	Multivariate A/B test (6)
Task				
Design and tune variants	33	22	13	2
Duration	45	9	11	2
Population assignment	37	7	8	2
Goal/hypothesis	27	17	8	2
Post-design actions	12	1	5	0
Pre-design actions	6	4	2	1
Metric analysis/init.	5	1	0	0

Table 11: Statistical methods \times A/B metrics (H short for hypothesis)

<i>Metric</i>	Engag.	Click	Monetary	Negative	Perf.	View	Feedback
<i>Method</i>							
H - equality	31	14	7	10	4	7	2
H - equality (unsp.)	24	12	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	4	1	2	1	0	1	0
H - independence	2	2	3	0	0	1	0
Regression method	1	1	1	0	0	1	0

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 [139], another study mentions determining the success condition of the A/B tests before executing them [151], and another study refers to providing an outcome range of the A/B tests [152].
- 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.

Table 12: Data collected for the A/B tests.

Data collected	Number of occurrences
Product/system data	48
User-centric data	26
Spatial-temporal data	20
Secondary data	6

- 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?

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.

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 (48 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.

Table 13: Evaluation method used in the primary studies.

Evaluation method	Number of occurrences
Empirical evaluation	100
Simulation based on real empirical data	26
Simulation	15
Illustrative example	10
Case study	5
Theoretical	1

Evaluation method. Table 13 summarizes the identified evaluation methods. The vast majority of primary studies provide results from an empirical evaluation (100 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 (26 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), and a single primary study provides a theoretical evaluation [121].

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 as 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 roles 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) [165, 28].

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 [125, 28] or a recommendation algorithm [65], and selecting a different visual design [8].	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 [165, 49].	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 [45], and further developing personalization methods [6].	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 [96], and designing a new A/B test to further test the quality of an A/B metric prediction model ¹⁰ [126].	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 [77], and the validation of a new statistical methodology for continuous monitoring of A/B tests [82].	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 [19].	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 [105].	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 [151].	4
Unspecified	The use of the test results was not specified in the study.	24

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 [151, 43].	19
	Experiment alteration	Altering aspects of the A/B test during its execution. Examples include adjusting the population assignment of the experiment [33], or adjusting the A/B variants themselves [152].	12
	Experiment termination	Stopping A/B tests when deemed necessary. Examples include manually stopping A/B tests when sufficient data is collected [96], or stopping the experiment early when harm is observed [89].	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 [71], performing a deeper analysis of suspicious results [60], or performing bias reduction techniques on the retrieved data from the A/B tests [110].	17
	Experiment triggering	Starting the execution of (subsequent) A/B tests [165, 22].	13
	Other	This category encompasses a few niche tasks, such as documenting the findings and learning from conducting the A/B test [144], rerunning A/B tests [112], or incorporating user feedback in the analysis of the A/B tests [106].	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”).

<i>Use</i>	<i>Task</i>	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	

- 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 [76]).
- 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) [57], and trigger the experiments (i.e. launch an experiment explicitly to fix a known bug in the application) [105].

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.
- 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 [78].
- 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.
- The only stakeholder task reported for theoretical evaluation is experiment alteration (primary study [121]).

Table 17: Evaluation method \times Tasks of stakeholders in the test execution ("emp. sim." short for "simulation based on real empirical data", "ill." short for "illustrative").

<i>Task Method</i>	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

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	34
Process-related	Add process guidelines	9
	Automate process	7
Quality-related	Enhance scalability	7
	Enhance applicability	6

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 (**D18**).

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. [45] 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. [15] 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 [126]. 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 [114], 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. [72]. 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. [83], 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. [73] 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 due 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. [71] 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. [57]. 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. [118] 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. [49] 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. [179]. 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. [153] 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 [111]. 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 [109] 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.

Table 19: Research topics of primary studies.

Topic	Number of occurrences	Primary studies
Application of A/B testing	51	[175, 123, 95, 102, 16, 15, 121, 171, 99, 148, 33, 70, 66, 52, 20, 174, 107, 63, 155, 150, 65, 163, 170, 27, 143, 2, 5, 135, 149, 7, 98, 147, 141, 173, 19, 26, 8, 114, 6, 122, 50, 97, 136, 125, 22, 124, 128, 159, 67, 3, 176]
Improving efficiency of A/B testing	20	[1, 28, 127, 164, 23, 85, 47, 39, 44, 86, 40, 45, 46, 109, 100, 83, 18, 78, 37, 64]
Beyond standard A/B testing	18	[166, 38, 82, 48, 77, 79, 112, 134, 72, 139, 126, 29, 118, 75, 49, 117, 30, 151]
Concrete A/B testing problems	17	[138, 73, 168, 105, 146, 43, 162, 14, 103, 111, 71, 24, 153, 137, 96, 101, 25]
Pitfalls and challenges of A/B testing	13	[91, 88, 54, 60, 167, 42, 169, 120, 11, 41, 110, 140, 90]
Experimentation frameworks and platforms	13	[144, 154, 106, 156, 108, 9, 131, 74, 21, 36, 179, 177, 152]
A/B testing at scale	9	[89, 160, 58, 165, 81, 157, 76, 57, 56]

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 141 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, etc¹¹).

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 [44, 127, 164, 85], investigating sequential testing techniques to stop A/B tests as soon as reasonable [83, 86, 1], proposing techniques to detect invalid A/B tests¹² [28], and using extra data such as periodicity patterns in user behavior to improve A/B testing [45].

¹¹See data item *A/B target* in Section 4.2 for specific references.

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

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

Environment	Number of occurrences
In-house experimentation system	20
Research tool or prototype	13
Commercial A/B testing tool	10
Commercial non A/B testing tool	7
User survey	1

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 [166, 48, 112], use of counterfactuals in the evaluation of A/B tests¹³ [77, 134], investigating ways of automating parts of the A/B testing process [139, 118, 117, 151], improving or altering the A/A testing process [79, 29], and investigating ways of combining offline- and online A/B testing [72, 126].

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 [96, 71], network A/B testing or A/B testing in marketplaces [103, 73, 24], A/B testing in the CPS domain with digital twins [43], or A/B testing for mobile applications [101, 168].

5.1.5. Pitfalls and challenges of A/B testing

This topic is about pitfalls related to conducting A/B testing [54, 88, 42, 41], or (particular domain-related) challenges related to A/B testing [169, 120, 110].

5.1.6. Experimentation frameworks and platforms

This topic covers papers that present an A/B testing platform [106, 152, 144], or a framework concerning aspects related to the A/B testing process such as a framework for detecting data loss in A/B tests [74], a framework for the design of A/B tests [36], or a framework for personalization of A/B testing [154].

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 [81, 157, 76], process models or guidelines for A/B testing at scale [58, 165], or concrete scalable solutions such as a scalable statistical method for measuring quantile treatment effects for performance metrics in A/B tests [160].

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 (20 occurrences), for instance dedicated environments developed by companies such as Microsoft [105], Google [152], eBay [157], and Etsy [83]. 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 [100]. Next, we observe research tools and prototypes (13 occurrences). Examples include a

¹³Counterfactual 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.

tool to perform online cloud experimentation [153], a research prototype for A/B testing implemented in NodeJS [139], a tool for A/B testing with decision assistants [96], and a tool that enables automatic execution of multiple A/B tests [151]. The remaining environments we identified were commercial A/B testing tools (10 occurrences), e.g., Optimizely [122], and Google Analytics [22]; commercial tools not related to A/B testing (7 occurrences), e.g., Crazy egg [22], a heatmap tool used to design A/B variants, and using Yahoo Gemini (advertisement platform) to test different advertising strategies [114]; 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 [152]. 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 [54], or manually checking for possible effects of running A/B tests by analyzing the deployed A/B tests in the system [157]. 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. [44], enabling proactive prediction of user behavior in A/B tests based on historical data [45], and a deeper study of A/B test estimators to achieve better sensitivity as mentioned by Poyarkov et al. [127].

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. [85] put forward learning sensitive combinations of A/B metrics as a general open problem, and Duan et al. [48] 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. [58, 169, 55]. In light of expanding this experimentation culture, (partial) automation of the A/B testing process is essential to enable and empower continuous experimentation [28, 71]. Initial research on automation of steps in the A/B testing has been conducted, as for example presented by Tamburrelli et al [151] and Mattos et al. [118], see Sections 5.1.3 and 4.5.2. 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. 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 [71], and automatically generating insights related to the rationale and cause of experiment results to experiment developers to guide product development [169].

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 [154, 2, 71], 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 [81, 14]. 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 [51]. However, one big downside of bootstrapping is that it is computationally expensive [110]. 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 [86], 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. [154] 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 used¹⁴ to obtain the results in the studies.

5.4. Threats to validity

In this section we list potential threats to the validity of the systematic literature review [10].

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 not all relevant works are covered. To mitigate this threat, we searched all main digital library sources that publish work in computer science. Secondly, we defined the search string by including all commonly used terms for A/B testing to ensure proper retrieval of relevant works. Lastly, we also applied snowballing on the selected papers from the automatic search query 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.

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

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.

Acknowledgement

We thank Michiel Provoost for his support to this study.

References

- [1] Vineet Abhishek and Shie Mannor. 2017. A Nonparametric Sequential Test for Online Randomized Experiments. In *Proceedings of the 26th International Conference on World Wide Web Companion* (Perth, Australia) (*WWW ’17 Companion*). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 610–616. <https://doi.org/10.1145/3041021.3054196>
- [2] Deepak Agarwal, Bo Long, Jonathan Traupman, Doris Xin, and Liang Zhang. 2014. LASER: A Scalable Response Prediction Platform for Online Advertising. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining* (New York, New York, USA) (*WSDM ’14*). Association for Computing Machinery, New York, NY, USA, 173–182. <https://doi.org/10.1145/2556195.2556252>
- [3] Michal Aharon, Yohay Kaplan, Rina Levy, Oren Somekh, Ayelet Blanc, Neetai Eshel, Avi Shahar, Assaf Singer, and Alex Zlotnik. 2019. 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* (Beijing, China) (*CIKM ’19*). Association for Computing Machinery, New York, NY, USA, 2793–2801. <https://doi.org/10.1145/3357384.3357801>
- [4] Michal Aharon, Oren Somekh, Avi Shahar, Assaf Singer, Baruch Trayvas, Hadas Vogel, and Dobri Dobrev. 2019. Carousel Ads Optimization in Yahoo Gemini Native. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) (*KDD ’19*). Association for Computing Machinery, New York, NY, USA, 1993–2001. <https://doi.org/10.1145/3292500.3330740>
- [5] Luca Aiello, Ioannis Arapakis, Ricardo Baeza-Yates, Xiao Bai, Nicola Barbieri, Amin Mantrach, and Fabrizio Silvestri. 2016. The Role of Relevance in Sponsored Search. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management* (Indianapolis, Indiana, USA) (*CIKM ’16*). Association for Computing Machinery, New York, NY, USA, 185–194. <https://doi.org/10.1145/2983323.2983840>

- [6] Ryuya Akase, Hiroto Kawabata, Akiomi Nishida, Yuki Tanaka, and Tamaki Kaminaga. 2021. Related Entity Expansion and Ranking Using Knowledge Graph. In *Complex, Intelligent and Software Intensive Systems*, Leonard Barolli, Kangbin Yim, and Tomoya Enokido (Eds.). Springer International Publishing, Cham, 172–184.
- [7] Rafael Alfaro-Flores, José Salas-Bonilla, Loic Juillard, and Juan Esquivel-Rodríguez. 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)*. 1–9. <https://doi.org/10.1109/CLEI53233.2021.9639977>
- [8] Joana Almeida and Beatriz Casais. 2022. Subject Line Personalization Techniques and Their Influence in the E-Mail Marketing Open Rate. In *Information Systems and Technologies*, Alvaro Rocha, Hojjat Adeli, Gintautas Dzemyda, and Fernando Moreira (Eds.). Springer International Publishing, Cham, 532–540.
- [9] Xavier Amatriain. 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* (San Francisco, California, USA) (CIKM ’13). Association for Computing Machinery, New York, NY, USA, 2201–2208. <https://doi.org/10.1145/2505515.2514701>
- [10] Apostolos Ampatzoglou, Stamatia Bibi, Paris Avgeriou, Marijn Verbeek, and Alexander Chatzigeorgiou. 2019. Identifying, categorizing and mitigating threats to validity in software engineering secondary studies. *Information and Software Technology* 106 (2019), 201–230. <https://doi.org/10.1016/j.infsof.2018.10.006>
- [11] Nirupama Appikata, Miao Chen, Michael Natkovich, and Joshua Walters. 2017. Demystifying dark matter for online experimentation. In *2017 IEEE International Conference on Big Data (Big Data)*. 1620–1626. <https://doi.org/10.1109/BigData.2017.8258096>
- [12] F. Auer and M. Felderer. 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, 335–344. <https://doi.org/10.1109/SEAA.2018.00062>
- [13] Florian Auer, Rasmus Ros, Lukas Kaltenbrunner, Per Runeson, and Michael Felderer. 2021. Controlled experimentation in continuous experimentation: Knowledge and challenges. *Information and Software Technology* 134 (2021), 106551. <https://doi.org/10.1016/j.infsof.2021.106551>
- [14] Eytan Bakshy and Eitan Frachtenberg. 2015. Design and Analysis of Benchmarking Experiments for Distributed Internet Services. In *Proceedings of the 24th International Conference on World Wide Web (Florence, Italy) (WWW ’15)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 108–118. <https://doi.org/10.1145/2736277.2741082>
- [15] Joel Barajas, Jaimie Kwon, Ram Akella, Aaron Flores, Marius Holtan, and Victor Andrei. 2012. Marketing Campaign Evaluation in Targeted Display Advertising. In *Proceedings of the Sixth International Workshop on Data Mining for Online Advertising and Internet Economy* (Beijing, China) (ADKDD ’12). Association for Computing Machinery, New York, NY, USA, Article 5, 7 pages. <https://doi.org/10.1145/2351356.2351361>
- [16] Joel Barajas, Jaimie Kwon, Ram Akella, Aaron Flores, Marius Holtan, and Victor Andrei. 2012. 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* (Beijing, China) (ADKDD ’12). Association for Computing Machinery, New York, NY, USA, Article 8, 9 pages. <https://doi.org/10.1145/2351356.2351364>
- [17] Victor R. Basili, Gianluigi Caldiera, and Dieter H. Rombach. 1994. *The Goal Question Metric Approach*. Vol. I. John Wiley & Sons.
- [18] Tobias 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. In *2013 International Conference on e-Business (ICE-B)*. 1–8.
- [19] Tobias Blask, Burkhardt Funk, and Reinhard Schulte. 2011. Should companies bid on their own brand in sponsored search?. In *Proceedings of the International Conference on e-Business*. 1–8.
- [20] Fedor Borisuk, Siddarth Malreddy, Jun Mei, Yiqun Liu, Xiaoyi Liu, Piyush Maheshwari, Anthony Bell, and Kaushik Rangadurai. 2021. VisRel: Media Search at Scale. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (Virtual Event, Singapore) (KDD ’21). Association for Computing Machinery, New York, NY, USA, 2584–2592. <https://doi.org/10.1145/3447548.3467081>
- [21] Slava Borodovsky and Saharon Rosset. 2011. A/B Testing at SweetIM: The Importance of Proper Statistical Analysis. In *2011 IEEE 11th International Conference on Data Mining Workshops*. 733–740. <https://doi.org/10.1109/ICDMW.2011.19>
- [22] Alex Brown, Binky Lush, and Bernard J. Jansen. 2016. Pixel efficiency analysis: A quantitative web analytics approach. *Proceedings of the Association for Information Science and Technology* 53, 1 (2016), 1–10. [https://doi.org/10.1002/pra2.2016.14505301040 arXiv:<https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/pra2.2016.14505301040>](https://doi.org/10.1002/pra2.2016.14505301040)
- [23] Roman Budylin, Alexey Drutsa, Ilya Katsev, and Valeriya Tsoy. 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* (Marina Del Rey, CA, USA) (WSDM ’18). Association for Computing Machinery, New York, NY, USA, 55–63. <https://doi.org/10.1145/3159652.3159699>
- [24] Tianchi Cai, Daxi Cheng, Chen Liang, Ziqi Liu, Lihong Gu, Huizhi Xie, Zhiqiang Zhang, Xiaodong Zeng, and Jinjie Gu. 2021. LinkLouvain: Link-Aware A/B Testing and Its Application on Online Marketing Campaign. In *Database Systems for Advanced Applications*, Christian S. Jensen, Ee-Peng Lim, De-Nian Yang, Wang-Chien Lee, Vincent S. Tseng, Vana Kalogeraki, Jen-Wei Huang, and Chih-Ya Shen (Eds.). Springer International Publishing, Cham, 499–510.
- [25] Javier Cámarra and Alfred Kobsa. 2009. Facilitating Controlled Tests of Website Design Changes: A Systematic Approach. In *Web Engineering*, Martin Gaedke, Michael Grossniklaus, and Oscar Díaz (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 370–378.

- [26] Samit Chakraborty, Md. Saiful Hoque, Naimur Rahman Jeem, Manik Chandra Biswas, Deepayan Bardhan, and Edgar Lobaton. 2021. Fashion Recommendation Systems, Models and Methods: A Review. *Informatics* 8, 3 (2021). <https://doi.org/10.3390/informatics8030049>
- [27] Guangde Chen, Bee-Chung Chen, and Deepak Agarwal. 2017. Social Incentive Optimization in Online Social Networks. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining* (Cambridge, United Kingdom) (*WSDM '17*). Association for Computing Machinery, New York, NY, USA, 547–556. <https://doi.org/10.1145/3018661.3018700>
- [28] Nanyu Chen, Min Liu, and Ya Xu. 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* (Melbourne VIC, Australia) (*WSDM '19*). Association for Computing Machinery, New York, NY, USA, 501–509. <https://doi.org/10.1145/3289600.3291000>
- [29] Russell Chen, Miao Chen, Mahendrasinh Ramsinh Jadav, Joonsuk Bae, and Don Matheson. 2017. Faster online experimentation by eliminating traditional A/A validation. In *2017 IEEE International Conference on Big Data (Big Data)*. 1635–1641. <https://doi.org/10.1109/BigData.2017.8258098>
- [30] Emmanuelle Claeys, Pierre Gançarski, Myriam Maumy-Bertrand, and Hubert Wassner. 2017. Regression Tree for Bandits Models in A/B Testing. In *Advances in Intelligent Data Analysis XVI*, Niall Adams, Allan Tucker, and David Weston (Eds.). Springer International Publishing, Cham, 52–62.
- [31] Rafael Costa, Elie Chéniaux, Pedro Rosaes, Marcele Carvalho, Rafael Freire, Márcio Versiani, Bernard Range, and Antonio Nardi. 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 (06 2011), 144–9. <https://doi.org/10.1590/S1516-44462011000200009>
- [32] John Creswell and Timothy Gutterman. 2018. *Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research*, 6th Edition. Pearson, New York, NY, USA.
- [33] Xinyi Dai, Yunjia Xi, Weinan Zhang, Qing Liu, Ruiming Tang, Xiuqiang He, Jiawei Hou, Jun Wang, and Yong Yu. 2021. Beyond Relevance Ranking: A General Graph Matching Framework for Utility-Oriented Learning to Rank. *ACM Trans. Inf. Syst.* 40, 2, Article 25 (nov 2021), 29 pages. <https://doi.org/10.1145/3464303>
- [34] Maya Daneva, Daniela Damian, Alessandro Marchetto, and Oscar Pastor. 2014. Empirical research methodologies and studies in Requirements Engineering: How far did we come? *Journal of Systems and Software* 95 (2014), 1–9. <https://doi.org/10.1016/j.jss.2014.06.035>
- [35] Rico de Feijter, Rob van Vliet, Erik Jagroep, Sietse Overbeek, and Sjaak Brinkkemper. 2017. *Towards the adoption of DevOps in software product organizations: A maturity model approach*. Technical Report. Utrecht University.
- [36] Wagner S. De Souza, Fernando O. Pereira, Vanessa G. Albuquerque, Jorge Melegati, and Eduardo Guerra. 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)*. 860–865. <https://doi.org/10.1109/COMPSAC54236.2022.00136>
- [37] Alex Deng. 2015. Objective Bayesian Two Sample Hypothesis Testing for Online Controlled Experiments. In *Proceedings of the 24th International Conference on World Wide Web* (Florence, Italy) (*WWW '15 Companion*). Association for Computing Machinery, New York, NY, USA, 923–928. <https://doi.org/10.1145/2740908.2742563>
- [38] Alex Deng, Tianxi Li, and Yu Guo. 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* (Seoul, Korea) (*WWW '14*). Association for Computing Machinery, New York, NY, USA, 609–618. <https://doi.org/10.1145/2566486.2568028>
- [39] Alex Deng, Yicheng Li, Jannan Lu, and Vivek Ramamurthy. 2021. On Post-Selection Inference in A/B Testing. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (Virtual Event, Singapore) (*KDD '21*). Association for Computing Machinery, New York, NY, USA, 2743–2752. <https://doi.org/10.1145/3447548.3467129>
- [40] Drew Dimmery, Eytan Bakshy, and Jasjeet Sekhon. 2019. Shrinkage Estimators in Online Experiments. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) (*KDD '19*). Association for Computing Machinery, New York, NY, USA, 2914–2922. <https://doi.org/10.1145/3292500.3330771>
- [41] Pavel Dmitriev, Brian Frasca, Somit Gupta, Ron Kohavi, and Garnet Vaz. 2016. Pitfalls of long-term online controlled experiments. In *2016 IEEE International Conference on Big Data (Big Data)*. 1367–1376. <https://doi.org/10.1109/BigData.2016.7840744>
- [42] Pavel Dmitriev, Somit Gupta, Dong Woo Kim, and Garnet Vaz. 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* (Halifax, NS, Canada) (*KDD '17*). Association for Computing Machinery, New York, NY, USA, 1427–1436. <https://doi.org/10.1145/3097983.3098024>
- [43] Jürgen Dobaj, Andreas Riel, Thomas Krug, Matthias Seidl, Georg Macher, and Markus Egretzberger. 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* (Pittsburgh, Pennsylvania) (*SEAMS '22*). Association for Computing Machinery, New York, NY, USA, 132–143. <https://doi.org/10.1145/3524844.3528057>
- [44] Alexey Drutsa, Gleb Gusev, and Pavel Serdyukov. 2015. 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* (Florence, Italy) (*WWW '15*). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 256–266. <https://doi.org/10.1145/2736277.2741116>

- [45] Alexey Drutsa, Gleb Gusev, and Pavel Serdyukov. 2017. Periodicity in User Engagement with a Search Engine and Its Application to Online Controlled Experiments. *ACM Trans. Web* 11, 2, Article 9 (apr 2017), 35 pages. <https://doi.org/10.1145/2856822>
- [46] Alexey Drutsa, Gleb Gusev, and Pavel Serdyukov. 2017. 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* (Perth, Australia) (*WWW '17*). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1301–1310. <https://doi.org/10.1145/3038912.3052664>
- [47] Alexey Drutsa, Anna Ufliand, and Gleb Gusev. 2015. Practical Aspects of Sensitivity in Online Experimentation with User Engagement Metrics. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management* (Melbourne, Australia) (*CIKM '15*). Association for Computing Machinery, New York, NY, USA, 763–772. <https://doi.org/10.1145/2806416.2806496>
- [48] Weitao Duan, Shan Ba, and Chunzhe Zhang. 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* (Virtual Event, Israel) (*WSDM '21*). Association for Computing Machinery, New York, NY, USA, 193–201. <https://doi.org/10.1145/3437963.3441737>
- [49] Wouter Duivesteijn, Tara Farzami, Thijs Putman, Evertjan Peer, Hilde J. P. Weerts, Jasper N. Adegeest, Gerson Foks, and Mykola Pechenizkiy. 2017. Have It Both Ways—From A/B Testing to A&B Testing with Exceptional Model Mining. In *Machine Learning and Knowledge Discovery in Databases*, Yasemin Altun, Kamalika Das, Taneli Mielikäinen, Donato Malerba, Jerzy Stefanowski, Jesse Read, Marinka Žitnik, Michelangelo Ceci, and Sašo Džeroski (Eds.). Springer International Publishing, Cham, 114–126.
- [50] Joshua Eckroth and Eric 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. *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* (2019), 9357 – 9364. <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85090801224&partnerID=40&md5=f3391d595e00df8a0cba7802c9043ebd> Cited by: 2.
- [51] B. Efron and R.J. Tibshirani. 1994. *An Introduction to the Bootstrap*. CRC Press. <https://books.google.be/books?id=MWC1DwAAQBAJ>
- [52] Beyza Ermis, Patrick Ernst, Yannik Stein, and Giovanni Zappella. 2020. Learning to Rank in the Position Based Model with Bandit Feedback. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (Virtual Event, Ireland) (*CIKM '20*). Association for Computing Machinery, New York, NY, USA, 2405–2412. <https://doi.org/10.1145/3340531.3412723>
- [53] Vladimir M. Erthal, Bruno P. de Souza, Paulo Sérgio M. dos Santos, and Guilherme H. Travassos. 2022. A Literature Study to Characterize Continuous Experimentation in Software Engineering. *CIBSE 2022 - XXV Ibero-American Conference on Software Engineering* (2022). <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85137064966&partnerID=40&md5=04240b73ab90eb841083173be558b33f> Cited by: 0.
- [54] Maria Esteller-Cucala, Vicenc Fernandez, and Diego Villuendas. 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* (Larnaca, Cyprus) (*UMAP'19 Adjunct*). Association for Computing Machinery, New York, NY, USA, 153–159. <https://doi.org/10.1145/3314183.3323853>
- [55] Aleksander Fabijan, Benjamin Arai, Pavel Dmitriev, and Lukas Vermeer. 2021. It takes a Flywheel to Fly: Kickstarting and Growing the A/B testing Momentum at Scale. In *2021 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*. 109–118. <https://doi.org/10.1109/SEAA53835.2021.00023>
- [56] Aleksander Fabijan, Pavel Dmitriev, Colin McFarland, Lukas Vermeer, Helena Holmström Olsson, and Jan Bosch. 2018. Experimentation growth: Evolving trustworthy A/B testing capabilities in online software companies. *Journal of Software: Evolution and Process* 30, 12 (2018), e2113. <https://doi.org/10.1002/smrv.2113> arXiv:<https://onlinelibrary.wiley.com/doi/pdf/10.1002/smrv.2113> e2113 JSME-17-0210.R2.
- [57] Aleksander Fabijan, Pavel Dmitriev, Helena Holmström Olsson, and Jan Bosch. 2017. The Benefits of Controlled Experimentation at Scale. In *2017 43rd Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*. 18–26. <https://doi.org/10.1109/SEAA.2017.47>
- [58] Aleksander Fabijan, Pavel Dmitriev, Helena Holmström Olsson, and Jan Bosch. 2017. The Evolution of Continuous Experimentation in Software Product Development: From Data to a Data-Driven Organization at Scale. In *Proceedings of the 39th International Conference on Software Engineering* (Buenos Aires, Argentina) (*ICSE '17*). IEEE Press, Los Alamitos, CA, USA, 770–780. <https://doi.org/10.1109/ICSE.2017.76>
- [59] A. Fabijan, P. Dmitriev, H. Holmstrom Olsson, and J. Bosch. 2020. The Online Controlled Experiment Lifecycle. *IEEE Software* 37, 02 (mar 2020), 60–67. <https://doi.org/10.1109/MS.2018.2875842>
- [60] Aleksander Fabijan, Jayant Gupchup, Somit Gupta, Jeff Omhover, Wen Qin, Lukas Vermeer, and Pavel Dmitriev. 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* (Anchorage, AK, USA) (*KDD '19*). Association for Computing Machinery, New York, NY, USA, 2156–2164. <https://doi.org/10.1145/3292500.3330722>
- [61] Aleksander Fabijan, Helena Holmström Olsson, and Jan Bosch. 2015. Customer Feedback and Data Collection Techniques in Software R&D: A Literature Review. In *Software Business*, João M. Fernandes, Ricardo J. Machado, and Krzysztof Wnuk (Eds.). Springer International Publishing, Cham, 139–153. https://doi.org/10.1007/978-3-319-19593-3_12
- [62] Aleksander Fabijan, Helena Holmström Olsson, and Jan Bosch. 2016. The Lack of Sharing of Customer Data in Large Software Organizations: Challenges and Implications. In *Agile Processes, in Software Engineering, and Ex-*

- treme Programming*, Helen Sharp and Tracy Hall (Eds.). Springer International Publishing, Cham, 39–52. https://doi.org/10.1007/978-3-319-33515-5_4
- [63] Yaron Fairstein, Elad Haramaty, Arnon Lazerson, and Liane Lewin-Eytan. 2022. External Evaluation of Ranking Models under Extreme Position-Bias. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining* (Virtual Event, AZ, USA) (*WSDM ’22*). Association for Computing Machinery, New York, NY, USA, 252–261. <https://doi.org/10.1145/3488560.3498420>
- [64] Elea McDonnell Feit and Ron Berman. 2019. Test & Roll: Profit-Maximizing A/B Tests. *Marketing Science* 38, 6 (2019), 1038–1058. <https://doi.org/10.1287/mksc.2019.1194>
- [65] Antonino Freno. 2017. Practical Lessons from Developing a Large-Scale Recommender System at Zalando. In *Proceedings of the Eleventh ACM Conference on Recommender Systems* (Como, Italy) (*RecSys ’17*). Association for Computing Machinery, New York, NY, USA, 251–259. <https://doi.org/10.1145/3109859.3109897>
- [66] Kun Fu, Fanlin Meng, Jieping Ye, and Zheng Wang. 2020. CompactETA: A Fast Inference System for Travel Time Prediction. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Virtual Event, CA, USA) (*KDD ’20*). Association for Computing Machinery, New York, NY, USA, 3337–3345. <https://doi.org/10.1145/3394486.3403386>
- [67] Burkhardt Funk. 2009. Optimizing price levels in e-commerce applications: An empirical study. *ICETE 2009 - International Joint Conference on e-Business and Telecommunications* (2009), 37 – 43. <https://www.scopus.com/inward/record.uri?eid=2-s2.0-74549181430&partnerID=40&md5=6dfdde67b807b3964c62fc8c1929dcf0> Cited by: 1.
- [68] Matthias Galster and Danny Weyns. 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, 11–20. <https://doi.org/10.1109/WICSA.2016.10>
- [69] Federico Giaimo, Hugo Andrade, and Christian Berger. 2020. Continuous experimentation and the cyber–physical systems challenge: An overview of the literature and the industrial perspective. *Journal of Systems and Software* 170 (2020), 110781. <https://doi.org/10.1016/j.jss.2020.110781>
- [70] Carlos A. Gomez-Uribe and Neil Hunt. 2016. The Netflix Recommender System: Algorithms, Business Value, and Innovation. *ACM Trans. Manage. Inf. Syst.* 6, 4, Article 13 (dec 2016), 19 pages. <https://doi.org/10.1145/2843948>
- [71] Anjan Goswami, Wei Han, Zhenrui Wang, and Angela Jiang. 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, 1094–1102. <https://doi.org/10.1109/BigData.2015.7363863>
- [72] Alois Gruson, Praveen Chandar, Christophe Charbuillet, James McInerney, Samantha Hansen, Damien Tardieu, and Ben Carterette. 2019. Offline Evaluation to Make Decisions About PlaylistRecommendation Algorithms. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (Melbourne VIC, Australia) (*WSDM ’19*). Association for Computing Machinery, New York, NY, USA, 420–428. <https://doi.org/10.1145/3289600.3291027>
- [73] Huan Gui, Ya Xu, Anmol Bhasin, and Jiawei Han. 2015. Network A/B Testing: From Sampling to Estimation. In *Proceedings of the 24th International Conference on World Wide Web* (Florence, Italy) (*WWW ’15*). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 399–409. <https://doi.org/10.1145/2736277.2741081>
- [74] Jayant Gupchup, Yasaman Hosseinkashi, Pavel Dmitriev, Daniel Schneider, Ross Cutler, Andrei Jefremov, and Martin Ellis. 2018. Trustworthy Experimentation Under Telemetry Loss. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (Torino, Italy) (*CIKM ’18*). Association for Computing Machinery, New York, NY, USA, 387–396. <https://doi.org/10.1145/3269206.3271747>
- [75] Shubham Gupta and Sneha Chokshi. 2020. Digital Marketing Effectiveness Using Incrementality. In *Advances in Computing and Data Sciences*, Mayank Singh, P. K. Gupta, Vipin Tyagi, Jan Flusser, Tuncer Ören, and Gianluca Valentino (Eds.). Springer Singapore, Singapore, 66–75.
- [76] Somit Gupta, Lucy Ulanova, Sumit Bhardwaj, Pavel Dmitriev, Paul Raff, and Aleksander Fabijan. 2018. The Anatomy of a Large-Scale Experimentation Platform. In *2018 IEEE International Conference on Software Architecture (ICSA)*. 1–109. <https://doi.org/10.1109/ICSA.2018.00009>
- [77] Viet Ha-Thuc, Avishek Dutta, Ren Mao, Matthew Wood, and Yunli Liu. 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* (Virtual Event, China) (*SIGIR ’20*). Association for Computing Machinery, New York, NY, USA, 2288–2296. <https://doi.org/10.1145/3397271.3401434>
- [78] Yan He and Miao Chen. 2017. A Probabilistic, Mechanism-Independent Outlier Detection Method for Online Experimentation. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. 640–647. <https://doi.org/10.1109/DSAA.2017.64>
- [79] Yan He, Lin Yu, Miao Chen, William Choi, and Don Matheson. 2022. A Cluster-Based Nearest Neighbor Matching Algorithm for Enhanced A/A Validation in Online Experimentation. In *Companion Proceedings of the Web Conference 2022* (Virtual Event, Lyon, France) (*WWW ’22*). Association for Computing Machinery, New York, NY, USA, 136–140. <https://doi.org/10.1145/3487553.3524220>
- [80] Jez Humble and David Farley. 2010. *Continuous Delivery: Reliable Software Releases through Build, Test, and Deployment Automation* (1st ed.). Addison-Wesley Professional, Illinois, IL, USA.
- [81] Hao Jiang, Fan Yang, and Wutao Wei. 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)*. 1533–1544. <https://doi.org/10.1109/BigData50022.2020.9377996>
- [82] Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. 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* (Halifax, NS, Canada) (*KDD '17*). Association for Computing Machinery, New York, NY, USA, 1517–1525. <https://doi.org/10.1145/3097983.3097992>
- [83] Nianqiao Ju, Diane Hu, Adam Henderson, and Liangjie Hong. 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* (Melbourne VIC, Australia) (*WSDM '19*). Association for Computing Machinery, New York, NY, USA, 492–500. <https://doi.org/10.1145/3289600.3291025>
- [84] Staffs Keele et al. 2007. *Guidelines for performing systematic literature reviews in software engineering*. Technical Report. Technical report, Ver. 2.3 EBSE Technical Report. EBSE.
- [85] Eugene Kharitonov, Alexey Drutsa, and Pavel Serdyukov. 2017. Learning Sensitive Combinations of A/B Test Metrics. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining* (Cambridge, United Kingdom) (*WSDM '17*). Association for Computing Machinery, New York, NY, USA, 651–659. <https://doi.org/10.1145/3018661.3018708>
- [86] Eugene Kharitonov, Aleksandr Vorobev, Craig Macdonald, Pavel Serdyukov, and Iadh Ounis. 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* (Santiago, Chile) (*SIGIR '15*). Association for Computing Machinery, New York, NY, USA, 473–482. <https://doi.org/10.1145/2766462.2767729>
- [87] Rochelle King, Elizabeth F. Churchill, and Caitlin Tan. 2017. *Designing with Data: Improving the User Experience with A/B Testing*. O'Reilly Media, Inc., Sebastopol, CA, USA.
- [88] Ron Kohavi, Alex Deng, Brian Frasca, Roger Longbotham, Toby Walker, and Ya Xu. 2012. Trustworthy Online Controlled Experiments: Five Puzzling Outcomes Explained. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Beijing, China) (*KDD '12*). Association for Computing Machinery, New York, NY, USA, 786–794. <https://doi.org/10.1145/2339530.2339653>
- [89] Ron Kohavi, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann. 2013. Online Controlled Experiments at Large Scale. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Chicago, Illinois, USA) (*KDD '13*). Association for Computing Machinery, New York, NY, USA, 1168–1176. <https://doi.org/10.1145/2487575.2488217>
- [90] Ron Kohavi, Alex Deng, Roger Longbotham, and Ya Xu. 2014. Seven Rules of Thumb for Web Site Experimenters. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, New York, USA) (*KDD '14*). Association for Computing Machinery, New York, NY, USA, 1857–1866. <https://doi.org/10.1145/2623330.2623341>
- [91] Ron Kohavi and Roger Longbotham. 2011. Unexpected Results in Online Controlled Experiments. *SIGKDD Explor. Newsl.* 12, 2 (mar 2011), 31–35. <https://doi.org/10.1145/1964897.1964905>
- [92] Ron Kohavi and Roger Longbotham. 2017. Online Controlled Experiments and A/B Testing. *Encyclopedia of machine learning and data mining* 7, 8 (2017), 922–929.
- [93] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal Henne. 2009. Controlled experiments on the web: Survey and practical guide. *Data Mining and Knowledge Discovery* 18 (02 2009), 140–181. <https://doi.org/10.1007/s10618-008-0114-1>
- [94] Ron Kohavi, Diane Tang, and Ya Xu. 2020. *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*. Cambridge University Press, Cambridge, United Kingdom. <https://doi.org/10.1017/9781108653985>
- [95] Anastasiia Kornilova and Lucas Bernardi. 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* (Virtual Event, Israel) (*WSDM '21*). Association for Computing Machinery, New York, NY, USA, 976–983. <https://doi.org/10.1145/3437963.3441812>
- [96] Kostantinos Koukouvis, Roberto Alcañiz Cubero, and Patrizio Pelliccione. 2016. A/B Testing in E-commerce Sales Processes. In *Software Engineering for Resilient Systems*, Ivica Crnkovic and Elena Troubitsyna (Eds.). Springer International Publishing, Cham, 133–148.
- [97] Anuj Kumar and Kartik Hosanagar. 2017. Measuring the Value of Recommendation Links on Product Demand. *SSRN Electronic Journal* (01 2017). <https://doi.org/10.2139/ssrn.2909971>
- [98] Ratnakar Kumar and Nitasha Hasteer. 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)*. 350–354. <https://doi.org/10.1109/CESYS.2017.8321296>
- [99] Mounia Lalmas, Janette Lehmann, Guy Shaked, Fabrizio Silvestri, and Gabriele Tolomei. 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* (Sydney, NSW, Australia) (*KDD '15*). Association for Computing Machinery, New York, NY, USA, 1929–1938. <https://doi.org/10.1145/2783258.2788581>
- [100] Minyong R. Lee and Milan Shen. 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* (London, United Kingdom) (*KDD '18*). Association for Computing Machinery, New York, NY, USA, 491–499. <https://doi.org/10.1145/3219819.3219905>
- [101] Florian Lettner, Clemens Holzmann, and Patrick Hufflesz. 2013. Enabling A/B Testing of Native Mobile Applications by Remote User Interface Exchange. In *Computer Aided Systems Theory - EUROCAST 2013*, Roberto Moreno-Díaz, Franz Pichler, and Alexis Quesada-Arencibia (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 458–466.
- [102] Chengbo Li, Lin Zhu, Guangyuan Fu, Longzhi Du, Canhua Zhao, Tianlun Ma, Chang Ye, and Pei Lee. 2021. Learning to Bundle Proactively for On-Demand Meal Delivery. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (Virtual Event, Queensland, Australia) (*CIKM '21*). Association for Computing

- Machinery, New York, NY, USA, 3898–3905. <https://doi.org/10.1145/3459637.3481931>
- [103] Hannah Li, Geng Zhao, Ramesh Johari, and Gabriel Y. Weintraub. 2022. Interference, Bias, and Variance in Two-Sided Marketplace Experimentation: Guidance for Platforms. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) (*WWW ’22*). Association for Computing Machinery, New York, NY, USA, 182–192. <https://doi.org/10.1145/3485447.3512063>
- [104] Lihong Li, Jin Young Kim, and Imed Zitouni. 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, 37–46. <https://doi.org/10.1145/2684822.2685311>
- [105] Paul Luo Li, Xiaoyu Chai, Frederick Campbell, Jilong Liao, Neeraja Abburu, Minsuk Kang, Irina Niculescu, Greg Brake, Siddharth Patil, James Dooley, and Brandon Paddock. 2021. 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)*. 170–179. <https://doi.org/10.1109/ICSE-SEIP52600.2021.00026>
- [106] Paul Luo Li, Pavel Dmitriev, Huibin Mary Hu, Xiaoyu Chai, Zoran Dimov, Brandon Paddock, Ying Li, Alex Kirshenbaum, Irina Niculescu, and Taj Thoresen. 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)*. 21–30. <https://doi.org/10.1109/ICSE-SEIP.2019.00011>
- [107] Yiyang Li, Guanyu Tao, Weinan Zhang, Yong Yu, and Jun Wang. 2017. Content Recommendation by Noise Contrastive Transfer Learning of Feature Representation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (Singapore, Singapore) (*CIKM ’17*). Association for Computing Machinery, New York, NY, USA, 1657–1665. <https://doi.org/10.1145/3132847.3132855>
- [108] Ye Li, Hong Xie, Yishi Lin, and John C.S. Lui. 2021. Unifying Offline Causal Inference and Online Bandit Learning for Data Driven Decision. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) (*WWW ’21*). Association for Computing Machinery, New York, NY, USA, 2291–2303. <https://doi.org/10.1145/3442381.3449982>
- [109] Kevin Liou and Sean J. Taylor. 2020. Variance-Weighted Estimators to Improve Sensitivity in Online Experiments. In *Proceedings of the 21st ACM Conference on Economics and Computation* (Virtual Event, Hungary) (*EC ’20*). Association for Computing Machinery, New York, NY, USA, 837–850. <https://doi.org/10.1145/3391403.3399542>
- [110] Sophia Liu, Aleksander Fabijan, Michael Furchtgott, Somit Gupta, Paweł Janowski, Wen Qin, and Pavel Dmitriev. 2019. Enterprise-Level Controlled Experiments at Scale: Challenges and Solutions. In *2019 45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*. 29–37. <https://doi.org/10.1109/SEAA.2019.00013>
- [111] Yuchu Liu, David Issa Mattos, Jan Bosch, Helena Holmström Olsson, and Jonn Lantz. 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)*. 300–307. <https://doi.org/10.1109/SEAA53835.2021.00046>
- [112] Widad Machmouchi, Ahmed Hassan Awadallah, Imed Zitouni, and Georg Buscher. 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* (Singapore, Singapore) (*CIKM ’17*). Association for Computing Machinery, New York, NY, USA, 757–765. <https://doi.org/10.1145/3132847.3132850>
- [113] Lech Madeyski, Wojciech Orzeszyna, Richard Torkar, and Mariusz Józala. 2014. Overcoming the Equivalent Mutant Problem: A Systematic Literature Review and a Comparative Experiment of Second Order Mutation. *IEEE Transactions on Software Engineering* 40, 1 (2014), 23–42. <https://doi.org/10.1109/TSE.2013.44>
- [114] Maria Madlberger and Jiri Jizdny. 2021. Impact of promotional social media content on click-through rate - Evidence from a FMCG company. *20th International Conferences on WWW/Internet 2021 and Applied Computing 2021* (2021), 3 – 10. <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85124068035&partnerID=40&md5=c0b8f49a3b48b3d561fd0ed305eb1895> Cited by: 0.
- [115] Sara Mahdavi-Hezavehi, Vinicius H.S. Durelli, Danny Weyns, and Paris Avgeriou. 2017. A systematic literature review on methods that handle multiple quality attributes in architecture-based self-adaptive systems. *Information and Software Technology* 90 (2017), 1–26. <https://doi.org/10.1016/j.infsof.2017.03.013>
- [116] Taisei Masuda, Kyoko Murakami, Kenkichi Sugiura, Sho Sakui, Ron Philip Schuring, and Mitsuhiro Mori. 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 (2022), 2044–2052. <https://doi.org/10.1016/j.vaccine.2022.02.030>
- [117] David Issa Mattos, Jan Bosch, and Helena Holmström Olsson. 2017. More for Less: Automated Experimentation in Software-Intensive Systems. In *Product-Focused Software Process Improvement*, Michael Felderer, Daniel Méndez Fernández, Burak Turhan, Marcos Kalinowski, Federica Sarro, and Dietmar Winkler (Eds.). Springer International Publishing, Cham, 146–161.
- [118] David Issa Mattos, Jan Bosch, and Helena Holmström Olsson. 2017. 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)*. 256–265. <https://doi.org/10.1109/SEAA.2017.15>
- [119] David Issa Mattos, Jan Bosch, and Helena Holmström Olsson. 2018. Challenges and Strategies for Undertaking Continuous Experimentation to Embedded Systems: Industry and Research Perspectives. In *Agile Processes in Software Engineering and Extreme Programming*, Juan Garbajosa, Xiaofeng Wang, and Ademar Aguiar (Eds.). Springer International Publishing, Cham, 277–292. https://doi.org/10.1007/978-3-319-91602-6_20
- [120] David Issa Mattos, Jan Bosch, Helena Holmstrom Olsson, Aita Maryam Korshani, and Jonn Lantz. 2020. Automotive A/B testing: Challenges and Lessons Learned from Practice. In *2020 46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*. 101–109. <https://doi.org/10.1109/SEAA51224.2020.00026>
- [121] Pavel Metrikov, Fernando Diaz, Sébastien Lahaie, and Justin Rao. 2014. Whole Page Optimization: How Page Elements

- Interact with the Position Auction. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation* (Palo Alto, California, USA) (*EC '14*). Association for Computing Machinery, New York, NY, USA, 583–600. <https://doi.org/10.1145/2600057.2602871>
- [122] Risto Miikulainen, Myles Brundage, Jonathan Epstein, Tyler Foster, Babak Hodjat, Neil Iscoe, Jingbo Jiang, Diego Legrand, Sam Nazari, Xin Qiu, Michael Scharff, Cory Schoolland, Robert Severn, and Aaron Shagrin. 2020. Ascend by Evolv: AI-Based Massively Multivariate Conversion Rate Optimization. *AI Magazine* 41, 1 (Apr. 2020), 44–60. <https://doi.org/10.1609/aimag.v41i1.5256>
- [123] Tadashi Okoshi, Kota Tsubouchi, and Hideyuki Tokuda. 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* (Anchorage, AK, USA) (*KDD '19*). Association for Computing Machinery, New York, NY, USA, 2792–2800. <https://doi.org/10.1145/3292500.3330732>
- [124] Takumi Ozawa, Akiyuki Sekiguchi, and Kazuhiko Tsuda. 2016. A Method for the Construction of User Targeting Knowledge for B2B Industry Website. *Procedia Computer Science* 96 (2016), 1147–1155. <https://doi.org/10.1016/j.procs.2016.08.157> Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 20th International Conference KES-2016.
- [125] Dan Pelleg, Oleg Rokhlenko, Idan Szpektor, Eugene Agichtein, and Ido Guy. 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* (San Francisco, California, USA) (*CSCW '16*). Association for Computing Machinery, New York, NY, USA, 1080–1090. <https://doi.org/10.1145/2818048.2820022>
- [126] Ladislav Peska and Peter Vojtas. 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* (Virtual Event, USA) (*HT '20*). Association for Computing Machinery, New York, NY, USA, 291–300. <https://doi.org/10.1145/3372923.3404781>
- [127] Alexey Poyarkov, Alexey Drutsa, Andrey Khalyavin, Gleb Gusev, and Pavel Serdyukov. 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* (San Francisco, California, USA) (*KDD '16*). Association for Computing Machinery, New York, NY, USA, 235–244. <https://doi.org/10.1145/2939672.2939688>
- [128] Jia Qu and Jing Zhang. 2016. Validating Mobile Designs with Agile Testing in China: Based on Baidu Map for Mobile. In *Design, User Experience, and Usability: Design Thinking and Methods*, Aaron Marcus (Ed.). Springer International Publishing, Cham, 491–498.
- [129] Federico Quin, Danny Weyns, and Matthias Galster. 2023. Study Systematic Literature Review on A/B Testing. https://people.cs.kuleuven.be/danny.weyns/material/SLR_AB/
- [130] Jan Renz, Daniel Hoffmann, Thomas Staibitz, and Christoph Meinel. 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, 304–313. <https://doi.org/10.1145/2883851.2883876>
- [131] Mohi Reza, Juho Kim, Ananya Bhattacharjee, Anna N. Rafferty, and Joseph Jay Williams. 2021. The MOOClet Framework: Unifying Experimentation, Dynamic Improvement, and Personalization in Online Courses. In *Proceedings of the Eighth ACM Conference on Learning @ Scale* (Virtual Event, Germany) (*L@S '21*). Association for Computing Machinery, New York, NY, USA, 15–26. <https://doi.org/10.1145/3430895.3460128>
- [132] Pilar Rodríguez, Alireza Haghaghkhah, Lucy Ellen Lwakatare, Susanna Teppola, Tanja Suomalainen, Juho Eskeli, Teemu Karvonen, Pasi Kuvaja, June M. Verner, and Markku Oivo. 2017. Continuous deployment of software intensive products and services: A systematic mapping study. *Journal of Systems and Software* 123 (2017), 263–291. <https://doi.org/10.1016/j.jss.2015.12.015>
- [133] Rasmus Ros and Per Runeson. 2018. Continuous Experimentation and A/B Testing: A Mapping Study. In *Proceedings of the 4th International Workshop on Rapid Continuous Software Engineering* (Gothenburg, Sweden) (*RCOSE '18*). Association for Computing Machinery, New York, NY, USA, 35–41. <https://doi.org/10.1145/3194760.3194766>
- [134] Nir Rosenfeld, Yishay Mansour, and Elad Yom-Tov. 2017. Predicting Counterfactuals from Large Historical Data and Small Randomized Trials. In *Proceedings of the 26th International Conference on World Wide Web Companion* (Perth, Australia) (*WWW '17 Companion*). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 602–609. <https://doi.org/10.1145/3041021.3054190>
- [135] Sandra Sajeev, Jade Huang, Nikos Karampatziakis, Matthew Hall, Sebastian Kochman, and Weizhu Chen. 2021. Contextual Bandit Applications in a Customer Support Bot. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (Virtual Event, Singapore) (*KDD '21*). Association for Computing Machinery, New York, NY, USA, 3522–3530. <https://doi.org/10.1145/3447548.3467165>
- [136] Suhrid Satyal, Ingo Weber, Hye-young Paik, Claudio Di Ciccio, and Jan Mendling. 2017. AB-BPM: Performance-Driven Instance Routing for Business Process Improvement. In *Business Process Management*, Josep Carmona, Gregor Engels, and Akhil Kumar (Eds.). Springer International Publishing, Cham, 113–129.
- [137] Suhrid Satyal, Ingo Weber, Hye young Paik, Claudio Di Ciccio, and Jan Mendling. 2019. Business process improvement with the AB-BPM methodology. *Information Systems* 84 (2019), 283–298. <https://doi.org/10.1016/j.is.2018.06.007>
- [138] Martin Saveski, Jean Pouget-Abadie, Guillaume Saint-Jacques, Weitao Duan, Souvik Ghosh, Ya Xu, and Edoardo M. Airoldi. 2017. Detecting Network Effects: Randomizing Over Randomized Experiments. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Halifax, NS, Canada) (*KDD '17*). Association for Computing Machinery, New York, NY, USA, 1027–1035. <https://doi.org/10.1145/3097983.3098192>
- [139] Gerald Schermann, Dominik Schöni, Philipp Leitner, and Harald C. Gall. 2016. Bifrost: Supporting Continuous De-

- ployment with Automated Enactment of Multi-Phase Live Testing Strategies. In *Proceedings of the 17th International Middleware Conference* (Trento, Italy) (*Middleware '16*). Association for Computing Machinery, New York, NY, USA, Article 12, 14 pages. <https://doi.org/10.1145/2988336.2988348>
- [140] Shahriar Shariat, Burkay Orten, and Ali Dasdan. 2017. Online Evaluation of Bid Prediction Models in a Large-Scale Computational Advertising Platform: Decision Making and Insights. *Knowl. Inf. Syst.* 51, 1 (apr 2017), 37–60. <https://doi.org/10.1007/s10115-016-0972-6>
- [141] Fanjuan Shi, Chirine Ghedira, and Jean-Luc Marini. 2015. Context Adaptation for Smart Recommender Systems. *IT Professional* 17, 6 (2015), 18–26. <https://doi.org/10.1109/MITP.2015.96>
- [142] Janet Siegmund, Norbert Siegmund, and Sven Apel. 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, 9–19. <https://doi.org/10.1109/ICSE.2015.24>
- [143] Natalia Silberstein, Oren Somekh, Yair Koren, Michal Aharon, Dror Porat, Avi Shahar, and Tingyi Wu. 2020. Ad Close Mitigation for Improved User Experience in Native Advertisements. In *Proceedings of the 13th International Conference on Web Search and Data Mining* (Houston, TX, USA) (*WSDM '20*). Association for Computing Machinery, New York, NY, USA, 546–554. <https://doi.org/10.1145/3336191.3371798>
- [144] Jorge Gabriel Siqueira and Melise M. V. de Paula. 2018. IPEAD A/B Test Execution Framework. In *Proceedings of the XIV Brazilian Symposium on Information Systems* (Caxias do Sul, Brazil) (*SBSI'18*). Association for Computing Machinery, New York, NY, USA, Article 14, 8 pages. <https://doi.org/10.1145/3229345.3229360>
- [145] Dan Siroker and Pete Koomen. 2013. *A/B Testing: The Most Powerful Way to Turn Clicks Into Customers* (1st ed.). Wiley Publishing, Hoboken, NJ, USA.
- [146] Bruce Spang, Veronica Hannan, Shravya Kunamalla, Te-Yuan Huang, Nick McKeown, and Ramesh Johari. 2021. Unbiased Experiments in Congested Networks. In *Proceedings of the 21st ACM Internet Measurement Conference (Virtual Event)* (*IMC '21*). Association for Computing Machinery, New York, NY, USA, 80–95. <https://doi.org/10.1145/3487552.3487851>
- [147] Akshitha Sriraman, Abhishek Dhanotia, and Thomas F. Wenisch. 2019. SoftSKU: Optimizing Server Architectures for Microservice Diversity @Scale. In *2019 ACM/IEEE 46th Annual International Symposium on Computer Architecture (ISCA)*. 513–526.
- [148] Fei Sun, Peng Jiang, Hanxiao Sun, Changhua Pei, Wenwu Ou, and Xiaobo Wang. 2018. Multi-Source Pointer Network for Product Title Summarization. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (Torino, Italy) (*CIKM '18*). Association for Computing Machinery, New York, NY, USA, 7–16. <https://doi.org/10.1145/3269206.3271722>
- [149] Idan Szpektor, Yoelle Maarek, and Dan Pelleg. 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* (Rio de Janeiro, Brazil) (*WWW '13*). Association for Computing Machinery, New York, NY, USA, 1249–1260. <https://doi.org/10.1145/2488388.2488497>
- [150] Yukihiko Tagami, Toru Hotta, Yusuke Tanaka, Shingo Ono, Koji Tsukamoto, and Akira Tajima. 2014. Filling Context-Ad Vocabulary Gaps with Click Logs. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, New York, USA) (*KDD '14*). Association for Computing Machinery, New York, NY, USA, 1955–1964. <https://doi.org/10.1145/2623330.2623334>
- [151] Giordano Tamburrelli and Alessandro Margara. 2014. Towards Automated A/B Testing. In *Search-Based Software Engineering*, Claire Le Goues and Shin Yoo (Eds.). Springer International Publishing, Cham, 184–198.
- [152] Diane Tang, Ashish Agarwal, Deirdre O'Brien, and Mike Meyer. 2010. Overlapping Experiment Infrastructure: More, Better, Faster Experimentation. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Washington, DC, USA) (*KDD '10*). Association for Computing Machinery, New York, NY, USA, 17–26. <https://doi.org/10.1145/1835804.1835810>
- [153] Mert Toslali, Srinivasan Parthasarathy, Fabio Oliveira, and Ayse K. Coskun. 2020. JACKPOT: Online experimentation of cloud microservices. *HotCloud 2020 - 12th USENIX Workshop on Hot Topics in Cloud Computing, co-located with USENIX ATC 2020* (2020). <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85091892156&partnerID=40&md5=cae12fe24f34f2b0818e448f8c07fb> Cited by: 1.
- [154] Ye Tu, Kinjal Basu, Cyrus DiCiccio, Romil Bansal, Preetam Nandy, Padmini Jaikumar, and Shaunak Chatterjee. 2021. Personalized Treatment Selection Using Causal Heterogeneity. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) (*WWW '21*). Association for Computing Machinery, New York, NY, USA, 1574–1585. <https://doi.org/10.1145/3442381.3450075>
- [155] Yutaro Ueoka, Kota Tsubouchi, and Nobuyuki Shimizu. 2020. Tackling Cannibalization Problems for Online Advertisement. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (Genoa, Italy) (*UMAP '20*). Association for Computing Machinery, New York, NY, USA, 358–362. <https://doi.org/10.1145/3340631.3394875>
- [156] Jean Vanderdonckt, Mathieu Zen, and Radu-Daniel Vatavu. 2019. AB4Web: An On-Line A/B Tester for Comparing User Interface Design Alternatives. *Proc. ACM Hum.-Comput. Interact.* 3, EICS, Article 18 (jun 2019), 28 pages. <https://doi.org/10.1145/3331160>
- [157] Deepak Kumar Vasthimal, Pavan Kumar Srirama, and Arun Kumar Akkinapalli. 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)*. 230–238. <https://doi.org/10.1109/BigDataSecurity-HPSC-IDS.2019.00052>
- [158] Daniel Walper, Julia Kassau, Philipp Methfessel, Timo Pronold, and Wolfgang Einhäuser. 2020. Optimizing user inter-

- faces 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 (ETRA '20 Short Papers)*. Association for Computing Machinery, New York, NY, USA, 1–4. <https://doi.org/10.1145/3379156.3391351>
- [159] Jian Wang and David Hardtke. 2015. User Latent Preference Model for Better Downside Management in Recommender Systems. In *Proceedings of the 24th International Conference on World Wide Web (Florence, Italy) (WWW '15)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1209–1219. <https://doi.org/10.1145/2736277.2741126>
- [160] Weinan Wang and Xi Zhang. 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 (Virtual Event, Israel) (WSDM '21)*. Association for Computing Machinery, New York, NY, USA, 202–210. <https://doi.org/10.1145/3437963.3441779>
- [161] Yu Wang, Somit Gupta, Jiannan Lu, Ali Mahmoudzadeh, and Sophia Liu. 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, 2425–2428. <https://doi.org/10.1145/3357384.3358143>
- [162] Zenan Wang, Carlos Carrion, Xiliang Lin, Fuhua Ji, Yongjun Bao, and Weipeng Yan. 2022. Adaptive Experimentation with Delayed Binary Feedback. In *Proceedings of the ACM Web Conference 2022 (Virtual Event, Lyon, France) (WWW '22)*. Association for Computing Machinery, New York, NY, USA, 2247–2255. <https://doi.org/10.1145/3485447.3512097>
- [163] Liang Wu and Mihajlo Grbovic. 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 (Virtual Event, China) (SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 2387–2396. <https://doi.org/10.1145/3397271.3401444>
- [164] Yuhang Wu, Zeyu Zheng, Guangyu Zhang, Zuohua Zhang, and Chu Wang. 2022. Non-Stationary A/B Tests. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (Washington DC, USA) (KDD '22)*. Association for Computing Machinery, New York, NY, USA, 2079–2089. <https://doi.org/10.1145/3534678.3539325>
- [165] Tong Xia, Sumit Bhardwaj, Pavel Dmitriev, and Aleksander Fabijan. 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)*. 11–20. <https://doi.org/10.1109/ICSE-SEIP.2019.00010>
- [166] Yuxiang Xie, Nanyu Chen, and Xiaolin Shi. 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 (London, United Kingdom) (KDD '18)*. Association for Computing Machinery, New York, NY, USA, 876–885. <https://doi.org/10.1145/3219819.3219860>
- [167] Yuxiang Xie, Meng Xu, Evan Chow, and Xiaolin Shi. 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 (Virtual Event, Israel) (WSDM '21)*. Association for Computing Machinery, New York, NY, USA, 949–957. <https://doi.org/10.1145/3437963.3441742>
- [168] Ya Xu and Nanyu Chen. 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 (San Francisco, California, USA) (KDD '16)*. Association for Computing Machinery, New York, NY, USA, 313–322. <https://doi.org/10.1145/2939672.2939703>
- [169] Ya Xu, Nanyu Chen, Addrian Fernandez, Omar Sinnou, and Anmol Bhasin. 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 (Sydney, NSW, Australia) (KDD '15)*. Association for Computing Machinery, New York, NY, USA, 2227–2236. <https://doi.org/10.1145/2783258.2788602>
- [170] Ye Xu, Zang Li, Abhishek Gupta, Ahmet Bugdayci, and Anmol Bhasin. 2014. Modeling Professional Similarity by Mining Professional Career Trajectories. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (New York, New York, USA) (KDD '14)*. Association for Computing Machinery, New York, NY, USA, 1945–1954. <https://doi.org/10.1145/2623330.2623368>
- [171] Yanbo Xu, Divyat Mahajan, Liz Manrao, Amit Sharma, and Emre Kiciman. 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 (Virtual Event, Israel) (WSDM '21)*. Association for Computing Machinery, New York, NY, USA, 409–417. <https://doi.org/10.1145/3437963.3441821>
- [172] Sezin Gizem Yaman, Myriam Munezero, Jürgen Münch, Fabian Fagerholm, Ossi Syd, Mika Aaltola, Christina Palmu, and Tomi Männistö. 2017. Introducing continuous experimentation in large software-intensive product and service organisations. *Journal of Systems and Software* 133 (2017), 195–211. <https://doi.org/10.1016/j.jss.2017.07.009>
- [173] Wanshan Yang, Gemeng Yang, Ting Huang, Lijun Chen, and Youjian Eugene Liu. 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)*. 4101–4108. <https://doi.org/10.1109/BigData.2018.8622067>
- [174] Runlong Ye, Pan Chen, Yini Mao, Angela Wang-Lin, Hammad Shaikh, Angela Zavaleta Bernuy, and Joseph Jay Williams. 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 (Chicago, IL, USA) (SIGITE '22)*. Association for Computing Machinery, New York, NY, USA, 16–22. <https://doi.org/10.1145/3537674.3554740>
- [175] Takeshi Yoneda, Shunsuke Kozawa, Keisuke Osone, Yukinori Koide, Yosuke Abe, and Yoshifumi Seki. 2019. Algorithms and System Architecture for Immediate Personalized News Recommendations. In *IEEE/WIC/ACM International Conference on Web Intelligence (Thessaloniki, Greece) (WI '19)*. Association for Computing Machinery, New York, NY,

- USA, 124–131. <https://doi.org/10.1145/3350546.3352509>
- [176] Scott W. H. Young. 2014. Improving Library User Experience with A/B Testing: Principles and Process. *Weave: Journal of Library User Experience* 1 (08 2014). <https://doi.org/10.3998/weave.12535642.0001.101>
- [177] Miao Yu, Wenbin Lu, and Rui Song. 2020. A new framework for online testing of heterogeneous treatment effect. *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence* (2020), 10310 – 10317. <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85106588123&partnerID=40&md5=53544f162212be7cd129e1f196debc8> Cited by: 2.
- [178] He Zhang and Muhammad Ali Babar. 2010. On Searching Relevant Studies in Software Engineering. In *Proceedings of the 14th International Conference on Evaluation and Assessment in Software Engineering (UK) (EASE'10)*. BCS Learning & Development Ltd., Swindon, GBR, 111–120.
- [179] Zhenyu Zhao, Yan He, and Miao Chen. 2017. Inform Product Change through Experimentation with Data-Driven Behavioral Segmentation. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. 69–78. <https://doi.org/10.1109/DSAA.2017.65>

Appendix A. List of primary studies

Table A.21: List of primary studies.

Paper ID	Reference	Title
1	[1]	A Nonparametric Sequential Test for Online Randomized Experiments
2	[138]	Detecting Network Effects: Randomizing Over Randomized Experiments
3	[91]	Unexpected Results in Online Controlled Experiments
4	[28]	How A/B Tests Could Go Wrong: Automatic Diagnosis of Invalid Online Experiments
5	[127]	Boosted Decision Tree Regression Adjustment for Variance Reduction in Online Controlled Experiments
6	[88]	Trustworthy Online Controlled Experiments: Five Puzzling Outcomes Explained
7	[89]	Online Controlled Experiments at Large Scale
8	[164]	Non-Stationary A/B Tests
9	[73]	Network A/B Testing: From Sampling to Estimation
10	[166]	False Discovery Rate Controlled Heterogeneous Treatment Effect Detection for Online Controlled Experiments
11	[54]	Experimentation Pitfalls to Avoid in A/B Testing for Online Personalization
12	[38]	Statistical Inference in Two-Stage Online Controlled Experiments with Treatment Selection and Validation
13	[23]	Consistent Transformation of Ratio Metrics for Efficient Online Controlled Experiments
14	[160]	CONQ: CONtinuous Quantile Treatment Effects for Large-Scale Online Controlled Experiments
15	[60]	Diagnosing Sample Ratio Mismatch in Online Controlled Experiments: A Taxonomy and Rules of Thumb for Practitioners
16	[144]	IPEAD A/B Test Execution Framework
17	[82]	Peeking at A/B Tests: Why It Matters, and What to Do about It
18	[85]	Learning Sensitive Combinations of A/B Test Metrics
19	[47]	Practical Aspects of Sensitivity in Online Experimentation with User Engagement Metrics
20	[168]	Evaluating Mobile Apps with A/B and Quasi A/B Tests
21	[39]	On Post-Selection Inference in A/B Testing
22	[48]	Online Experimentation with Surrogate Metrics: Guidelines and a Case Study
23	[44]	Future User Engagement Prediction and Its Application to Improve the Sensitivity of Online Experiments
24	[58]	The Evolution of Continuous Experimentation in Software Product Development: From Data to a Data-Driven Organization at Scale
25	[167]	How to Measure Your App: A Couple of Pitfalls and Remedies in Measuring App Performance in Online Controlled Experiments
26	[86]	Sequential Testing for Early Stopping of Online Experiments
27	[40]	Shrinkage Estimators in Online Experiments
28	[77]	A Counterfactual Framework for Seller-Side A/B Testing on Marketplaces
29	[45]	Periodicity in User Engagement with a Search Engine and Its Application to Online Controlled Experiments
30	[105]	Evolving Software to be ML-Driven Utilizing Real-World A/B Testing: Experiences, Insights, Challenges

31	[46]	Using the Delay in a Treatment Effect to Improve Sensitivity and Preserve Directionality of Engagement Metrics in A/B Experiments
32	[79]	A Cluster-Based Nearest Neighbor Matching Algorithm for Enhanced A/A Validation in Online Experimentation
33	[109]	Variance-Weighted Estimators to Improve Sensitivity in Online Experiments
34	[42]	A Dirty Dozen: Twelve Common Metric Interpretation Pitfalls in Online Controlled Experiments
35	[100]	Winner’s Curse: Bias Estimation for Total Effects of Features in Online Controlled Experiments
36	[169]	From Infrastructure to Culture: A/B Testing Challenges in Large Scale Social Networks
37	[83]	A Sequential Test for Selecting the Better Variant: Online A/B Testing, Adaptive Allocation, and Continuous Monitoring
38	[146]	Unbiased Experiments in Congested Networks
39	[154]	Personalized Treatment Selection Using Causal Heterogeneity
40	[112]	Beyond Success Rate: Utility as a Search Quality Metric for Online Experiments
41	[175]	Algorithms and System Architecture for Immediate Personalized News Recommendations
42	[106]	Experimentation in the Operating System: The Windows Experimentation Platform
43	[156]	AB4Web: An On-Line A/B Tester for Comparing User Interface Design Alternatives
44	[123]	Real-World Product Deployment of Adaptive Push Notification Scheduling on Smartphones
45	[95]	Mining the Stars: Learning Quality Ratings with User-Facing Explanations for Vacation Rentals
46	[43]	Towards Digital Twin-Enabled DevOps for CPS Providing Architecture-Based Service Adaptation & Verification at Runtime
47	[162]	Adaptive Experimentation with Delayed Binary Feedback
48	[108]	Unifying Offline Causal Inference and Online Bandit Learning for Data Driven Decision
49	[9]	Beyond Data: From User Information to Business Value through Personalized Recommendations and Consumer Science
50	[102]	Learning to Bundle Proactively for On-Demand Meal Delivery
51	[16]	Measuring Dynamic Effects of Display Advertising in the Absence of User Tracking Information
52	[15]	Marketing Campaign Evaluation in Targeted Display Advertising
53	[121]	Whole Page Optimization: How Page Elements Interact with the Position Auction
54	[131]	The MOOClet Framework: Unifying Experimentation, Dynamic Improvement, and Personalization in Online Courses
55	[171]	Split-Treatment Analysis to Rank Heterogeneous Causal Effects for Prospective Interventions
56	[99]	Promoting Positive Post-Click Experience for In-Stream Yahoo Gemini Users
57	[134]	Predicting Counterfactuals from Large Historical Data and Small Randomized Trials
58	[148]	Multi-Source Pointer Network for Product Title Summarization
59	[33]	Beyond Relevance Ranking: A General Graph Matching Framework for Utility-Oriented Learning to Rank
60	[72]	Offline Evaluation to Make Decisions About PlaylistRecommendation Algorithms

61	[74]	Trustworthy Experimentation Under Telemetry Loss
62	[70]	The Netflix Recommender System: Algorithms, Business Value, and Innovation
63	[139]	Bifrost: Supporting Continuous Deployment with Automated Enactment of Multi-Phase Live Testing Strategies
64	[66]	CompactETA: A Fast Inference System for Travel Time Prediction
65	[14]	Design and Analysis of Benchmarking Experiments for Distributed Internet Services
66	[52]	Learning to Rank in the Position Based Model with Bandit Feedback
67	[20]	VisRel: Media Search at Scale
68	[174]	Behavioral Consequences of Reminder Emails on Students' Academic Performance: A Real-World Deployment
69	[107]	Content Recommendation by Noise Contrastive Transfer Learning of Feature Representation
70	[63]	External Evaluation of Ranking Models under Extreme Position-Bias
71	[155]	Tackling Cannibalization Problems for Online Advertisement
72	[150]	Filling Context-Ad Vocabulary Gaps with Click Logs
73	[65]	Practical Lessons from Developing a Large-Scale Recommender System at Zalando
74	[163]	How Airbnb Tells You Will Enjoy Sunset Sailing in Barcelona? Recommendation in a Two-Sided Travel Marketplace
75	[170]	Modeling Professional Similarity by Mining Professional Career Trajectories
76	[27]	Social Incentive Optimization in Online Social Networks
77	[143]	Ad Close Mitigation for Improved User Experience in Native Advertisements
78	[126]	Off-Line vs. On-Line Evaluation of Recommender Systems in Small E-Commerce
79	[2]	LASER: A Scalable Response Prediction Platform for Online Advertising
80	[5]	The Role of Relevance in Sponsored Search
81	[135]	Contextual Bandit Applications in a Customer Support Bot
82	[165]	Safe Velocity: A Practical Guide to Software Deployment at Scale using Controlled Rollout
83	[149]	When Relevance is Not Enough: Promoting Diversity and Freshness in Personalized Question Recommendation
84	[103]	Interference, Bias, and Variance in Two-Sided Marketplace Experimentation: Guidance for Platforms
85	[120]	Automotive A/B testing: Challenges and Lessons Learned from Practice
86	[21]	A/B Testing at SweetIM: The Importance of Proper Statistical Analysis
87	[36]	A Framework Model to Support A/B Tests at the Class and Component Level
88	[81]	Statistical Reasoning of Zero-Inflated Right-Skewed User-Generated Big Data A/B Testing
89	[111]	Size matters? Or not: A/B testing with limited sample in automotive embedded software
90	[157]	Scalable Data Reporting Platform for A/B Tests
91	[18]	Applying Bayesian parameter estimation to A/B tests in e-business applications examining the impact of green marketing signals in sponsored search advertising
92	[7]	Experiment-driven improvements in Human-in-the-loop Machine Learning Annotation via significance-based A/B testing
93	[76]	The Anatomy of a Large-Scale Experimentation Platform
94	[11]	Demystifying dark matter for online experimentation
95	[71]	Controlled experiments for decision-making in e-Commerce search
96	[98]	Evaluating usability of a web application: A comparative analysis of open-source tools

97	[29]	Faster online experimentation by eliminating traditional A/A validation
98	[41]	Pitfalls of long-term online controlled experiments
99	[147]	SoftSKU: Optimizing Server Architectures for Microservice Diversity @Scale
100	[57]	The Benefits of Controlled Experimentation at Scale
101	[141]	Context Adaptation for Smart Recommender Systems
102	[173]	Whales, Dolphins, or Minnows? Towards the Player Clustering in Free Online Games Based on Purchasing Behavior via Data Mining Technique
103	[110]	Enterprise-Level Controlled Experiments at Scale: Challenges and Solutions
104	[19]	Should companies bid on their own brand in sponsored search?
105	[78]	A Probabilistic, Mechanism-Independent Outlier Detection Method for Online Experimentation
106	[179]	Inform Product Change through Experimentation with Data-Driven Behavioral Segmentation
107	[118]	Your System Gets Better Every Day You Use It: Towards Automated Continuous Experimentation
108	[26]	Fashion Recommendation Systems, Models and Methods: A Review
109	[8]	Subject Line Personalization Techniques and Their Influence in the E-Mail Marketing Open Rate
110	[114]	Impact of promotional social media content on click-through rate - Evidence from a FMCG company
111	[6]	Related Entity Expansion and Ranking Using Knowledge Graph
112	[24]	LinkLouvain: Link-Aware A/B Testing and Its Application on Online Marketing Campaign
113	[122]	Ascend by Evolv: Artificial intelligence-based massively multivariate conversion rate optimization
114	[177]	A new framework for online testing of heterogeneous treatment effect
115	[153]	JACKPOT: Online experimentation of cloud microservices
116	[75]	Digital Marketing Effectiveness Using Incrementality
117	[137]	Business process improvement with the AB-BPM methodology
118	[50]	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	[97]	Measuring the value of recommendation links on product demand
120	[56]	Experimentation growth: Evolving trustworthy A/B testing capabilities in online software companies
121	[140]	Online Evaluation of Bid Prediction Models in a Large-Scale Computational Advertising Platform: Decision Making and Insights
122	[136]	AB-BPM: Performance-driven instance routing for business process improvement
123	[49]	Have It Both Ways—From A/B Testing to A&B Testing with Exceptional Model Mining
124	[117]	More for Less: Automated Experimentation in Software-Intensive Systems
125	[30]	Regression Tree for Bandits Models in A/B Testing
126	[125]	When the Crowd is Not Enough: Improving User Experience with Social Media through Automatic Quality Analysis
127	[22]	Pixel efficiency analysis: A quantitative web analytics approach
128	[96]	A/B Testing in E-commerce Sales Processes
129	[124]	A Method for the Construction of User Targeting Knowledge for B2B Industry Website
130	[128]	Validating Mobile Designs with Agile Testing in China: Based on Baidu Map for Mobile
131	[159]	User Latent Preference Model for Better Downside Management in Recommender Systems

132	[151]	Towards Automated A/B Testing
133	[90]	Seven rules of thumb for web site experimenters
134	[101]	Enabling A/B Testing of Native Mobile Applications by Remote User Interface Exchange
135	[152]	Overlapping Experiment Infrastructure: More, Better, Faster Experimentation
136	[67]	Optimizing price levels in e-commerce applications: An empirical study
137	[25]	Facilitating Controlled Tests of Website Design Changes: A Systematic Approach
138	[3]	Soft Frequency Capping for Improved Ad Click Prediction in Yahoo Gemini Native
139	[37]	Objective Bayesian Two Sample Hypothesis Testing for Online Controlled Experiments
140	[64]	Test & Roll: Profit-Maximizing A/B Tests
141	[176]	Improving Library User Experience with A/B Testing: Principles and Process