Data Driven A or B Testing Methodology for Website Effectiveness

Qurotul Aini[1], Aulia Khanza[2], Vinkan Likita[3], Steven Harazaki Lase[4], Yasir Mustafa Kareem[5*]
[1]Faculty of Information Technology, Satya Wacana Christian University, Indonesia
[2,3,4]Faculty of Economics and Business, University of Raharja, Indonesia
[5]Department of Digital Business, Eesp Incorporation, British Indian Ocean Territory
[1]aini@raharja.info, [2]aulia.khanza@raharja.info, [3]vinkan@raharja.info, [4]steven.harazaki@raharja.info, [5]mustafa.kar33m@eesp.io
***Corresponding Author**

Article Info

Article history:

Submission January 14, 2026
Revised February 6, 2026
Accepted February 9, 2026

Keywords:

A or B Testing
Big Data Analytics
Website
Optimization
Data-Driven

ABSTRACT

Website design and optimization decisions are often driven by subjective opinions, internal organizational preferences, or prevailing industry trends rather than empirical evidence derived from large-scale user interaction data, resulting in suboptimal performance and inconsistent user experiences. In digital environments characterized by high data volume and velocity, the absence of a structured experimentation methodology limits organizations' ability to effectively leverage Big Data for continuous website improvement. **This paper presents** a comprehensive and systematic methodological guide to A or B testing as a data-driven approach for enhancing website effectiveness in data-intensive contexts. Unlike existing A or B testing guides that focus mainly on tools or isolated experimental outcomes, this study proposes an end-to-end framework integrating hypothesis formulation, scalable experimental design, statistical rigor, iterative learning, and practical decision-making into a unified and replicable process. **The methodology** outlines the complete A or B testing lifecycle, including alignment of business objectives with measurable data signals, development of testable hypotheses, controlled experiment implementation, large-scale data collection, and statistical analysis to ensure validity and significance of findings. **The results** demonstrate that a disciplined and continuous A or B testing program supported by Big Data analytics enables incremental yet compounding improvements in website performance. Through illustrative case examples, the study shows that relatively small, data-informed changes to website elements such as headlines, calls-to-action, images, and layout structures can lead to statistically significant gains in conversion rates, user engagement, and overall user experience. **The paper concludes** that A or B testing serves as a strategic Big Data analytics mechanism that supports evidence-based website optimization decisions grounded in empirical user behavior rather than intuition.

DOI: https://doi.org/10.33050/corisinta.v3n1.170

1. INTRODUCTION

The rapid growth of digital marketing has significantly transformed how organizations design, evaluate, and optimize their websites in increasingly data-intensive environments [1]. Decisions that were once driven by intuition or aesthetic preference are now expected to be supported by empirical evidence derived from large-scale user interaction data generated across digital platforms [2]. The emergence of accessible A or B testing tools, experimentation features embedded within analytics systems, and integrated conversion rate optimization solutions has expanded organizations' capacity to systematically test and refine website elements [3]. In Big Data contexts, these tools enable the collection and analysis of high-volume, high-velocity behavioral data, allowing organizations to move toward evidence-based optimization strategies. This shift reflects a broader paradigm of data-driven design, where optimization efforts are guided by statistically validated insights and measurable performance outcomes rather than subjective judgment [4].

Despite the widespread availability of experimentation technologies, many organizations continue to implement A or B testing in an ad-hoc and methodologically unstructured manner [5]. Experiments are often poorly designed, focus on low-impact variables, or are conducted with insufficient sample sizes and inadequate test durations, which is particularly problematic in Big Data environments [6]. Furthermore, the misinterpretation of large-scale experimental data without proper statistical controls frequently leads to false positives and unreliable conclusions [7]. These weaknesses reduce organizational trust in experimentation results and limit the strategic value of Big Data analytics for decision-making [8]. The absence of a rigorous and scalable experimentation framework ultimately constrains the effectiveness of A or B testing as a reliable mechanism for improving website performance and user experience at scale [9].

In response to these challenges, this study is guided by research questions that examine the formulation of valid and testable hypotheses, the design and execution of A or B tests in data-rich environments, and the interpretation of results in terms of both statistical significance and practical relevance [5]. The primary objective of this research is to propose a formal six-step A or B testing framework that integrates core statistical principles with scalable data collection, analysis, and learning processes suitable for Big Data applications [6, 10]. The scope of the study is limited to A or B testing on website landing pages, emphasizing incremental optimization rather than radical redesign, which typically requires alternative qualitative approaches [11]. This paper is structured to review relevant literature, present the proposed methodology, discuss illustrative case examples, and conclude with implications and future research directions aligned with Big Data analytics, applied experimentation, and computational decision-support systems [12].

The proposed framework is designed to be scalable in environments characterized by high data volume, velocity, and variety [13]. By aligning A or B testing with distributed data processing architectures, automated data collection mechanisms, and AI-driven analytics, the methodology enables experimentation to be conducted efficiently at scale [14]. Integration with machine learning models allows experimental results to inform adaptive decision systems, supporting continuous optimization in dynamic digital environments where user behavior evolves rapidly [15]. Within the scope of Big Data research, this study positions A or B testing not merely as a tactical optimization technique but as a scalable analytical mechanism capable of supporting data-driven decision-making in complex digital systems [16]. The proposed framework emphasizes the role of large-scale data processing, statistical reliability, and repeatable experimentation cycles as foundational components for extracting actionable insights from massive user interaction datasets [17]. By framing A or B testing as part of a broader Big Data analytics pipeline, this research contributes to the understanding of how controlled experiments can enhance system performance, support adaptive digital infrastructures, and strengthen evidence-based strategies in data-intensive web environments [18]. This perspective reinforces the relevance of the study to contemporary research domains encompassing Big Data analytics, applied computational experimentation, and data-driven system optimization [19].

In addition to its methodological and technical contributions, this study aligns with several Sustainable Development Goals by promoting more efficient, transparent, and responsible use of digital data in organizational decision-making [20]. The structured application of A or B testing supported by Big Data analytics contributes to SDGs 9 (Industry, Innovation, and Infrastructure) by enabling organizations to build more adaptive, data-driven, and innovative digital infrastructures [21]. Furthermore, by reducing trial-and-error decision-making and minimizing ineffective design changes, the proposed framework supports SDGs 12 (Responsible Consumption and Production) through more efficient use of computational resources, digital assets, and development efforts. The emphasis on evidence-based optimization and continuous learning also indirectly supports SDGs 8 (Decent Work and Economic Growth) by improving digital performance outcomes that enhance organizational productivity and competitiveness.

2. LITERATURE REVIEW

2.1. The Scientific Method in Business and Big Data Contexts

The application of the scientific method has increasingly expanded beyond traditional scientific disciplines into business and digital marketing, particularly within data intensive environments. In the context of Big Data, A or B testing represents a practical implementation of scientific reasoning applied to large scale user behavior analysis, where decisions are derived from empirical evidence generated by high volume and high velocity data rather than intuition or subjective judgment. The scientific method emphasizes hypothesis formulation, controlled experimentation, and objective evaluation of outcomes, which are essential for extracting reliable insights from complex digital datasets [22]. Within A or B testing, the original website version functions as the control, while the modified version represents the experimental condition, with user traffic distributed across both versions at scale [23]. By isolating a single variable in a data rich environment, organizations can attribute observed performance differences such as changes in conversion rates or engagement metrics directly to the tested design element with statistical confidence. Prior studies indicate that adopting a scientific and data driven mindset in business experimentation improves decision quality, reduces cognitive bias, and enables organizations to systematically leverage Big Data analytics for continuous optimization and evidence based digital strategy development [24].

2.2. The Foundations of Statistical Hypothesis Testing

Statistical hypothesis testing forms the analytical backbone of A or B testing. Central to this process is the null hypothesis, which assumes that no meaningful difference exists between the control and the variant design [25]. The p-value is then used to estimate the probability of obtaining the observed results if the null hypothesis were true. A result is typically considered statistically significant when the p-value falls below a predefined threshold, commonly $p < 0.05$, allowing researchers to reject the null hypothesis with reasonable confidence [26]. In addition to significance testing, statistical power plays a critical role by determining a test's ability to detect real effects when they exist. Insufficient sample sizes or prematurely terminated experiments reduce statistical power, increasing the risk of false conclusions. The literature consistently highlights that misunderstanding these statistical principles is a major source of error in practical A or B testing implementations.

2.3. Heuristic Frameworks for Conversion Optimization

To guide hypothesis generation, researchers and practitioners often rely on heuristic frameworks for conversion rate optimization. Models such as the LIFT framework which emphasizes value proposition, relevance, clarity, urgency, and anxiety reduction and the MECLABS conversion formula provide structured lenses for identifying potential optimization opportunities [27]. These frameworks help translate qualitative insights about user behavior into testable hypotheses, ensuring that A or B tests are strategically grounded rather than randomly selected. Prior research suggests that combining heuristic frameworks with rigorous statistical testing enhances both the efficiency and effectiveness of website optimization efforts.

3. METHODOLOGY

The A or B testing methodology proposed in this study consists of six main steps, systematically structured and designed to be compatible with Big Data environments and artificial intelligence based systems. These six steps form an end-to-end workflow that integrates data analysis, experiment design, technical implementation, and continuous learning [28]. This approach aims to ensure that experiments are not only statistically valid, but also practically relevant and scalable across complex digital systems.

3.1. Initial Analysis and Data Governance

The initial phase focuses on a thorough analysis of available quantitative and qualitative data to identify key issues and optimization opportunities. Quantitative data can include performance metrics such as conversion rates, bounce rates, visit times, and user navigation flows, while qualitative data is obtained through heatmaps, session recordings, or user feedback. In the context of Big Data, this phase also includes evaluating data quality, data governance, and consistency across data sources [29]. Data governance is crucial in large-scale experiments, as inconsistent metric definitions, missing data, or tracking bias can significantly impact the validity of the results. Therefore, establishing data standards, validation mechanisms, and quality control are necessary to ensure that the data used in experiments accurately represents user behavior.

3.2. Hypothesis Formulation

Based on the results of the initial analysis, a hypothesis is formulated explicitly by linking specific design changes to the performance metrics to be improved. A good hypothesis not only states the expected difference but also includes a rationale supported by historical data or user behavior theory. With this approach, each experiment has a clear purpose and can be objectively tested [30]. In a Big Data environment, structured hypothesis formulation helps prioritize experiments with the greatest potential impact. This is crucial to avoid experiments that are high in cost but low in information value [31].

3.3. Variation Design and Scalability

The variation design phase focuses on developing alternative versions of the elements being tested while maintaining the principle of variable isolation [32]. Design changes are limited to a single primary variable to facilitate attribution of experimental results. From a technical perspective, this design must support large amounts of distributed traffic and be compatible with parallel processing architectures and distributed systems [33]. Scalability is a key consideration, especially when experiments are run concurrently or on platforms with millions of users. Therefore, the design of variations must consider page load efficiency, consistency of user experience, and the system's ability to handle the experiment's load.

3.4. Experiment Implementation

Experiment implementation is conducted using a testing system capable of handling high-speed and large volumes of data. This process includes randomizing users into control and variation groups, automatically logging user interactions, and integrating with a Big Data pipeline for data collection and storage [34]. In the context of AI-based systems, experiment implementation can also be integrated with adaptive analytics modules that enable real-time performance monitoring. This approach helps ensure that the experiment is running according to design and detects anomalies early.

3.5. Analysis and Inference

After the experiment is completed, the data is analyzed using appropriate statistical methods to evaluate performance differences between the control and variation groups. The analysis includes statistical significance testing, calculating confidence intervals, and estimating effect sizes [35]. In a Big Data environment, the interpretation of results depends not only on statistical significance but also on practical relevance in the context of business objectives. A cautious approach to inference is necessary to avoid misleading conclusions, especially when small differences become statistically significant due to very large sample sizes.

3.6. Learning and Iteration

The final stage highlights continuous learning as a central component of effective A or B testing. Each experiment provides meaningful insights into user behavior, regardless of whether the results are statistically significant. Positive outcomes validate hypotheses and identify changes that improve performance, while non-significant results help eliminate low-impact variables and refine future testing priorities [36]. This learning process is inherently iterative, as insights from one experiment inform the design of subsequent tests. By systematically incorporating findings into new hypotheses and optimization strategies, organizations can gradually enhance website effectiveness and reduce uncertainty in decision-making. Over time, this iterative approach supports sustainable, data-driven improvement rather than isolated, one-off optimizations.

Table 1. Mapping of A or B Testing Methodological Steps to Data Outputs and Decisions

Stage	Input	Output	Decision
Initial Analysis	User data	Identified key issues	Define test focus
Test Design	Data and design changes	Hypotheses and variants	Run experiment
Execution and Analysis	Experimental data	Statistical results	Accept or reject hypothesis
Learning	Test insights	Updated knowledge	Plan next iteration

Table 1 summarizes the core stages of the proposed A or B testing methodology by condensing the experimentation process into four essential phases. The table illustrates how user interaction data are progressively transformed into actionable decisions through a structured and iterative workflow. The process begins with initial analysis, where user behavior data are examined to identify key performance issues and determine the focus of testing [37]. This stage ensures that experiments are problem-driven rather than exploratory or intuition-based. The test design stage translates analytical findings into testable hypotheses and clearly defined control and variant versions, establishing a rigorous experimental setup. The execution and analysis phase involves running the experiment, collecting behavioral data, and applying statistical methods to evaluate performance differences between variants. The outcomes of this stage inform evidence-based decisions regarding whether a design change should be accepted or rejected. Finally, the learning stage emphasizes the iterative nature of A or B testing, where insights from completed experiments are incorporated into organizational knowledge and used to guide subsequent optimization cycles [38].

4. RESULT AND DISCUSSION

This section presents illustrative case studies that demonstrate the application of the proposed A or B testing methodology in real-world website optimization. The cases translate the conceptual framework into practical experiments and include both successful and non-significant results, emphasizing that all outcomes provide valuable insights when properly interpreted [39]. By evaluating elements such as call-to-action buttons, headlines, and visual components, the case studies highlight the importance of hypothesis-driven testing, sufficient test duration, and rigorous statistical analysis in data-driven decision-making.

4.1. The Call-to-Action (CTA) Test

The first case study examines the effectiveness of CTA button text on a website landing page, a critical element in guiding user actions and influencing conversion outcomes [40]. In this experiment, the control version (Version A) employed a generic CTA label, “Submit”, which provided minimal information about the value users would receive after clicking. In contrast, the variation (Version B) used a more benefit-oriented phrase, “Get Your Free Guide”, explicitly communicating the perceived value of the action. The underlying hypothesis proposed that emphasizing value and clarity in CTA messaging would increase user motivation and reduce uncertainty, thereby improving conversion rates [41]. The experiment was designed following the proposed methodology, with random traffic allocation and a predefined sample size to ensure sufficient statistical power. After running the test for an adequate duration, the results indicated a statistically significant increase in conversions for Version B. These findings confirm that value-driven CTA wording can substantially influence user behavior by aligning messaging with user expectations and perceived benefits.

4.2. The Headline Test

The second case study focuses on headline optimization, another high-impact element in website design that plays a central role in shaping first impressions and user engagement. In the control condition (Version A), the landing page featured a feature-focused headline, “The Best Insurance Platform”, which emphasized product attributes without directly addressing user outcomes. The variation (Version B) adopted a benefit-oriented headline, “Save 20% on Your Insurance”, highlighting a tangible and immediately relevant user benefit. The hypothesis underlying this experiment suggested that benefit-oriented messaging would resonate more strongly with users by directly addressing their motivations and goals. The A or B test results showed that Version B outperformed Version A across key performance metrics, including engagement and conversion rates. This outcome reinforces existing research that users respond more positively to messaging that emphasizes personal value rather than abstract product superiority. The case underscores the importance of framing content in a way that aligns with user intent and decision-making processes.

4.3. A Failed Test and Its Learnings

The third case study presents an example of a failed A or B test, in which no statistically significant difference was observed between the control and variation. In this experiment, the tested change involved altering the color of a secondary button, with the hypothesis that increased visual contrast would draw more user attention and lead to higher conversions. Despite careful implementation and sufficient sample size, the experiment yielded minimal differences in performance metrics, resulting in the rejection of the hypothesis [42]. Rather than being interpreted as an unsuccessful outcome, this result provided valuable insight into user behavior and design priorities. The findings suggested that button color, at least in this context, was not a primary factor influencing user decisions. As a result, the experimentation team was able to redirect resources toward testing more impactful elements, such as value propositions, content hierarchy, and page layout. This case highlights a critical principle of data-driven experimentation. Failure tests are an essential component of the learning process and play a key role in refining future optimization strategies.

Table 2. Summary of A or B Testing Outcomes

Element Tested	Metric	Outcome
CTA Text	Conversion Rate	Significant improvement
Headline	Conversion Rate	Significant improvement
Button Color	Click-through Rate	No significant effect

Table 2 summarizes the main outcomes of the A or B testing experiments conducted in this study by presenting the tested website elements, the primary performance metrics, and the observed results. It provides a compact overview of how different types of design and content changes influence user behavior and conversion performance. By focusing on key metrics, the table allows for direct comparison across experiments and highlights patterns in experimental effectiveness. The results indicate that modifications to high-impact, message-driven elements specifically call-to-action text and headline messaging produced statistically significant improvements in conversion rates. These findings suggest that clearly communicating user value and benefits plays a more decisive role in influencing user decisions than purely aesthetic adjustments. In contrast, the experiment involving button color changes did not yield a significant effect on click-through rates, indicating that visual changes alone may have limited influence when not supported by meaningful informational cues. Overall, the table reinforces the role of A or B testing as a learning-oriented, evidence-based approach to website optimization. By presenting both significant and non-significant outcomes, it demonstrates that all experimental results contribute valuable insights, helping organizations prioritize impactful design decisions and avoid assumptions that are not supported by empirical data.

Figure 1. The Iterative Learning Cycle in A or B Testing

Figure 1 depicts the iterative learning cycle that forms the foundation of effective A or B testing as a data-driven experimentation approach. The process begins with the formulation of a hypothesis derived from prior analysis, user insights, or heuristic evaluation, followed by the execution of an A or B test in a controlled experimental environment [43]. After the experiment concludes, the results whether statistically significant or not are analyzed to evaluate differences in performance metrics and to ensure that conclusions are supported by empirical evidence. The figure also highlights how insights generated from each experiment inform subsequent iterations of testing and optimization. Positive results guide the implementation of validated changes, while non-significant outcomes contribute equally valuable knowledge by identifying elements with limited impact on user behavior. Through this continuous feedback loop, A or B testing enables systematic learning and incremental improvement, reinforcing its role as an ongoing process rather than a one-time optimization effort.

This discusses the broader implications of implementing A or B testing as a core decision-making tool rather than merely a tactical optimization technique. When adopted at a strategic level, A or B testing supports a systematic shift toward evidence-based decision-making by enabling organizations to evaluate design and content changes using empirical user data [44]. In this context, A or B testing functions not only as a method for improving isolated website elements but also as an analytical mechanism embedded within broader data-driven governance and digital performance management frameworks. Beyond methodological considerations, effective A or B testing requires a high level of organizational readiness. This includes the availability of reliable data infrastructure, clearly defined performance metrics, and disciplined experimental procedures. Equally important is the presence of an organizational mindset that values experimentation, learning, and continuous improvement. Without such alignment, even well-designed experiments risk being underutilized, misinterpreted, or overridden by subjective decision-making. Therefore, this discussion also highlights common implementation challenges and clarifies the practical boundaries of A or B testing, offering a balanced perspective on its role in improving website effectiveness within complex digital environments.

4.4. Building a Culture of Experimentation

Successful A or B testing requires a fundamental organizational shift toward a culture of continuous experimentation. In such a culture, experimentation is not treated as a one-time activity but as an ongoing process that informs design, marketing, and strategic decisions. Decision-making is guided by data and empirical evidence rather than intuition, personal preference, or organizational hierarchy, often referred to as the Highest Paid Person's Opinion (HiPPO) effect [44]. By reducing reliance on subjective judgment, organizations can achieve more objective, transparent, and reproducible optimization outcomes. A strong experimentation culture also encourages cross-functional collaboration among teams such as design, marketing, engineering, and analytics. These teams must work together to formulate hypotheses, design experiments, and interpret results within a shared analytical framework. Furthermore, organizations must foster an environment where unsuccessful or non-significant test results are viewed as valuable learning opportunities rather than failures. This perspective supports long-term capability building and enables organizations to accumulate institutional knowledge about user behavior, ultimately strengthening data-driven website optimization practices.

4.5. Common Pitfalls to Avoid

Despite its conceptual simplicity, A or B testing is susceptible to several common pitfalls that can significantly reduce the reliability and credibility of experimental results. One frequent mistake is terminating experiments prematurely before sufficient sample sizes or test durations are reached, which undermines statistical power and increases the risk of false conclusions [45]. Another issue is the misinterpretation of statistical significance, where minor differences are overemphasized without considering their practical or business relevance. Additional challenges arise when multiple variables are tested simultaneously without proper experimental controls, making it difficult to attribute observed effects to specific changes. In Big Data environments, these pitfalls can be amplified because very large sample sizes may produce statistically significant results that lack meaningful impact. Avoiding such errors requires disciplined experimental planning, adherence to statistical best practices, and clear communication of experimental assumptions and limitations. Addressing these challenges is essential to maintaining trust in the experimentation process and ensuring that A or B testing contributes reliably to decision-making.

4.6. The Limits of A or B Testing

While A or B testing is highly effective for achieving incremental performance improvements, it is important to recognize its inherent limitations, particularly when applied in complex digital environments [46]. The method is best suited for optimizing existing designs, interfaces, or messaging through small, controlled changes where causal relationships can be clearly isolated. However, in environments characterized by multivariable user journeys, dynamic content rendering, and personalized experiences, traditional A or B testing may face challenges related to interaction effects and experimental isolation. These conditions can complicate causal inference, making it difficult to attribute observed outcomes to a single design change. Moreover, A or B testing is not designed to generate entirely new value propositions, uncover latent user needs, or address fundamental product or usability issues. Such challenges typically require exploratory and qualitative research methods, including user interviews, usability testing, and design thinking approaches that extend beyond controlled experimentation [47, 48]. In contexts where system behavior is highly adaptive or influenced by multiple interacting variables, A or B testing should therefore be complemented with multivariate experimentation, causal modeling, or AI-based optimization techniques to better capture the complexity of user behavior and system dynamics. Acknowledging these limitations allows organizations to position A or B testing appropriately within a broader research and innovation ecosystem. When combined with complementary methods such as user research, multivariate analysis, and data-driven modeling, A or B testing can play a crucial role in continuous optimization while avoiding unrealistic expectations. This balanced perspective ensures that A or B testing remains a powerful yet properly scoped tool for improving website effectiveness and digital performance in increasingly complex digital systems.

4.7. A or B Testing within Big Data and AI-Driven Decision Systems

In Big Data and AI-driven environments, A or B testing should be understood as an integral component of a broader analytical workflow rather than as an isolated experimental activity. User interaction data are typically captured through large-scale logging systems and event streams, which are then processed using distributed analytics pipelines capable of handling high data volume and velocity [49]. Within such architectures, experimental data generated from A or B testing are integrated alongside behavioral, transactional, and contextual data to support comprehensive analysis and informed decision-making. Within this ecosystem, A or B testing functions as a critical validation layer that provides causal evidence to complement predictive insights produced by machine learning models. While AI-based systems can identify patterns, forecast user behavior, or automate personalization, A or B testing ensures that these predictions translate into measurable performance improvements when applied in real-world conditions [50]. By embedding experimentation directly into Big Data pipelines, organizations can maintain empirical accountability in automated or AI-driven decision processes, ensuring that optimization strategies remain aligned with observed user behavior and continuously adapt to changing digital environments.

5. MANAGERIAL IMPLICATIONS

This outlines the managerial implications of implementing A or B testing as a structured, data-driven methodology for website optimization in Big Data environments. Beyond its technical application, A or B testing has important strategic, organizational, and governance-related consequences for managers responsible for digital performance, analytics, and innovation. The implications discussed below emphasize how managerial practices, resource allocation, and organizational culture must evolve to fully leverage A or B testing as a reliable decision-support mechanism.

5.1. A or B Testing as a Strategic Decision-Support Tool

Managers should view A or B testing as a strategic analytical mechanism rather than a purely tactical optimization technique. When embedded within managerial decision-making processes, A or B testing enables leaders to evaluate design and content changes based on empirical evidence derived from large-scale user behavior data. This shift reduces the influence of intuition, hierarchy-driven decisions, and personal preferences, allowing managerial actions to be justified through statistically validated outcomes. As a result, strategic website decisions such as messaging, layout prioritization, and conversion optimization can be made with greater confidence and accountability.

5.2. Alignment of Experimentation with Business Objectives

An important managerial implication of this study is the need to align experimentation activities with clearly defined business objectives and key performance indicators. Managers play a critical role in ensuring that A or B testing initiatives focus on variables with meaningful business impact rather than low-value or purely cosmetic changes. By linking hypotheses directly to organizational goals such as conversion growth, user engagement, or retention, managers can prioritize experiments more effectively. This alignment also supports more efficient use of resources by directing time, budget, and development effort toward experiments that generate actionable insights and measurable value.

5.3. Organizational Readiness and Cross-Functional Coordination

Effective A or B testing requires organizational readiness that extends beyond technical infrastructure. Managers must facilitate collaboration across design, marketing, engineering, and analytics teams to ensure that experiments are properly designed, implemented, and interpreted. Clear roles, standardized metrics, and shared analytical frameworks help reduce miscommunication and inconsistent interpretations of results. From a managerial perspective, establishing governance structures and experimentation guidelines is essential for maintaining methodological rigor and ensuring that experimental findings are trusted and adopted across the organization.

5.4. Managing Risk, Learning, and Decision Accountability

The findings also highlight the managerial responsibility to manage risk and uncertainty in digital optimization initiatives. A or B testing allows managers to test changes in controlled environments, thereby reducing the risk of implementing ineffective or harmful design decisions at scale. Importantly, managers must foster a learning-oriented mindset in which non-significant or negative results are treated as valuable insights rather than failures. This approach strengthens decision accountability by ensuring that all outcomes contribute to organizational knowledge and inform future optimization strategies.

5.5. Implications for Big Data and AI-Driven Digital Management

In Big Data and AI-driven environments, managers must integrate A or B testing into broader analytical and decision-making ecosystems. Experimental results provide causal validation for insights generated by predictive analytics and machine learning models, helping managers distinguish correlation from causation. By embedding A or B testing within Big Data pipelines and AI-driven personalization systems, managers can ensure that automated or adaptive decisions remain empirically grounded. This integration supports more responsible, transparent, and sustainable digital management practices in increasingly complex web environments.

6. CONCLUSION

This study confirms that A or B testing, when implemented with methodological rigor and robust statistical principles, is an effective approach for achieving continuous, data-driven improvements in website performance. A structured experimentation process enables organizations to transition from intuition-based decision-making toward evidence-based optimization grounded in large-scale user behavior data. Through systematic stages ranging from hypothesis formulation to controlled experimentation and iterative learning, organizations can improve decision quality and achieve measurable gains in key performance indicators such as conversion rates and overall user experience, particularly within Big Data environments.

The primary contribution of this paper lies in the development of a practical and scalable A or B testing framework aligned with contemporary Big Data analytics and data-driven decision-making practices. Rather than treating A or B testing as a series of isolated experiments, the proposed framework integrates statistical foundations, heuristic optimization strategies, and real-world case illustrations into a unified methodological approach. By positioning A or B testing as a strategic analytical mechanism, this study supports sustainable experimentation practices and contributes to more efficient digital infrastructure development, in line with the objectives of SDGs 9 and SDGs 12.

Future research may extend this work by integrating artificial intelligence and machine learning techniques to enable adaptive experimentation, automated personalization, and more complex experimental designs such as A or B/n and multivariate testing. Further investigation is also required to address emerging challenges related to data privacy regulations, modern web architectures, and experimental validity in increasingly complex digital ecosystems. Addressing these issues will be essential to ensure that A or B testing remains a robust and relevant methodology for data-intensive systems in the future.

7. DECLARATIONS

7.1. About Authors

Qurotul Aini (QA) https://orcid.org/0000-0002-7546-5721

Aulia Khanza (AK) https://orcid.org/0009-0007-3736-4317

Vinkan Likita (VL) https://orcid.org/0009-0009-6435-8494

Steven Harazaki Lase (SH) https://orcid.org/0009-0002-6647-7541

Yasir Mustafa Kareem (YM) https://orcid.org/0009-0008-5096-2300

7.2. Author Contributions

Conceptualization: QA; Methodology: AK; Software: VL; Validation: SH and YM; Formal Analysis: QA and AK; Investigation: VL; Resources: SH; Data Curation: YM; Writing Original Draft Preparation: QA and AK; Writing Review and Editing: VL and SH; Visualization: YM; All authors, QA, AK, VL, SH and YM have read and agreed to the published version of the manuscript.

7.3. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

7.4. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

7.5. Declaration of Conflicting Interest

The authors declare that they have no conflicts of interest, known competing financial interests, or personal relationships that could have influenced the work reported in this paper.

REFERENCES

[1] Y. Zanichkovskyy, V. Fast, and A. Masyuk, "Adaptive cdn node selection in dynamic ict systems using online controlled experiments and change detection multi armed bandit algorithms," 2025.

[2] U. Rahardja, M. Miftah, M. Rakhmansyah, and J. Zanubiya, "Revolutionizing financial services with big data and fintech: A scalable approach to innovation," *ADI Journal on Recent Innovation*, vol. 6, no. 2, pp. 118–129, 2025.

[3] H. Rahman, M. T. Chowdhury, M. I. Sumon, F. M. Mim, M. R. A. Rashid, R. U. Islam, and M. Hasan, "Analyzing a/b testing results in digital marketing," in *2024 2nd International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings)*. IEEE, 2024, pp. 1–5.

[4] S. Kumar, "B testing in personalization: examining the effectiveness of personalized user experiences based on a/b test results," *J Artif Intell Mach Learn Data Sci*, vol. 1, no. 4, pp. 546–50, 2023.

[5] N. Larsen, J. Stallrich, S. Sengupta, A. Deng, R. Kohavi, and N. T. Stevens, "Statistical challenges in online controlled experiments: A review of a/b testing methodology," *The American Statistician*, vol. 78, no. 2, pp. 135–149, 2024.

[6] M. Mandić, I. Gregurec, and U. Vujović, "Measuring the effectiveness of online sales by conducting a/b testing," *Market-Tržište*, vol. 35, no. 2, pp. 223–249, 2023.

[7] D. Hernandez, L. Pasha, D. A. Yusuf, R. Nurfaizi, and D. Julianingsih, "The role of artificial intelligence in sustainable agriculture and waste management: Towards a green future," *International Transactions on Artificial Intelligence*, vol. 2, no. 2, pp. 150–157, 2024.

[8] A. B. Ahady, E.-M. Klopries, H. Schüttrumpf, and S. Wolf, "Drought analysis methods: a multidisciplinary review with insights on key decision-making factors in method selection," *Water*, vol. 17, no. 15, p. 2248, 2025.

[9] R. Koning, S. Hasan, and A. Chatterji, "Experimentation and start-up performance: Evidence from a/b testing," *Management Science*, vol. 68, no. 9, pp. 6434–6453, 2022.

[10] J. R. Saura, D. Palacios-Marqués, and D. Ribeiro-Soriano, "Digital marketing in smes via data-driven strategies: Reviewing the current state of research," *Journal of Small Business Management*, vol. 61, no. 3, pp. 1278–1313, 2023.

[11] K. Sun, L. Kong, H. Zhu, and C. Shi, "Arma-design: Optimal treatment allocation strategies for a/b testing in partially observable time series experiments," *arXiv preprint arXiv:2408.05342*, 2024.

[12] F. Xiong, C. He, Y. Liu, A. M. Black-Schaffer, and T. Nag, "Distinct quasiparticle interference patterns for surface impurity scattering on various weyl semimetals," *Physical Review B*, vol. 109, no. 5, p. 054201, 2024.

[13] G. Silva, G. Godwin, and O. Jayanagara, "The impact of ai on personalized learning and educational analytics," *International Transactions on Education Technology (ITEE)*, vol. 3, no. 1, pp. 36–46, 2024.

[14] Q. Zhang and L. Kang, "Locally optimal design for a/b tests in the presence of covariates and network dependence," *Technometrics*, vol. 64, no. 3, pp. 358–369, 2022.

[15] F. A. Buot, G. T. Maglasang, and A. R. B. Elnar, "Unification of mixed hilbert-space representations in condensed matter physics and quantum field theory," *arXiv preprint arXiv:2204.07691*, 2022.

[16] D. Garcia, J. Tolvanen, and A. K. Wagner, "Demand estimation using managerial responses to automated price recommendations," *Management Science*, vol. 68, no. 11, pp. 7918–7939, 2022.

[17] R. Johari, P. Koomen, L. Pekelis, and D. Walsh, "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*, 2017, pp. 1517–1525.

[18] Y. Liu, Y. Zhou, P. Li, and F. Hu, "Cluster-adaptive network a/b testing: From randomization to estimation," *Journal of Machine Learning Research*, vol. 25, no. 170, pp. 1–48, 2024.

[19] H. Xu, Y. Man, M. Yang, J. Wu, Q. Zhang, and J. Wang, "Analytical insight of earth: a cloud-platform of intelligent computing for geospatial big data," *arXiv preprint arXiv:2312.16385*, 2023.

[20] N. Larsen, J. Stallrich, S. Sengupta, A. Deng, R. Kohavi, and N. T. Stevens, "Supplemental material for "statistical challenges in online controlled experiments: A review of a/b testing methodology"."

[21] X. Wu, J. C.-A. Tsai, and Y. Lei, "Information technology project management research: A review of works by influential pioneers," *Project Management Journal*, vol. 54, no. 4, pp. 366–391, 2023.

[22] D. Martinez, L. Magdalena, and A. N. Savitri, "Ai and blockchain integration: Enhancing security and transparency in financial transactions," *International Transactions on Artificial Intelligence*, vol. 3, no. 1, pp. 11–20, 2024.

[23] Z. Sun, X. Wang, T. Han, J. Liu, L. Wang, X. Wei, H. Huang, X. Huang, and Z. Wu, "Advanced prediction of pipeline vertical deformation and axial strain via multi-source data fusion and multi-task deep learning," *Structural Health Monitoring*, p. 14759217251385078, 2025.

[24] V. Garg and P. Singh, "Optimizing digital flyer experiences with data integration for e-commerce," *Integr. J. Res. Arts Humanit*, 2024.

[25] N. Lutfiani, N. P. L. Santoso, R. Ahsanitaqwim, U. Rahardja, and A. R. A. Zahra, "Ai-based strategies to improve resource efficiency in urban infrastructure," *International Transactions on Artificial Intelligence*, vol. 2, no. 2, pp. 121–127, 2024.

[26] V. Sharma, M. Gupta, A. K. Pandey, D. Mishra, and A. Kumar, "A review of deep learning-based human activity recognition on benchmark video datasets," *Applied Artificial Intelligence*, vol. 36, no. 1, p. 2093705, 2022.

[27] T. P. Vo, M. T. Nguyen, T. A. Nguyen, G. V. Nguyen, H. N. Nguyen, G. B. Pham, M. H. Ha, and Q. D. Nguyen, "Improving the recovery of phenolics and flavonoids from coriandrum sativum l. leaves using microwave-enzymatic-assisted extraction," *Sustainable Food Technology*, 2025.

[28] M. Ahli, M. F. Hilmi, and A. Abudaqa, "Ethical sales behavior influencing trust, loyalty, green experience, and satisfaction in uae public entrepreneur firms," *Aptisi Transactions on Technopreneurship (ATT)*, vol. 6, no. 2, pp. 149–168, 2024.

[29] Y. Ma, L. Liu, Z. Liu, F. Li, Q. Xie, K. Chen, C. Lv, Y. He, and F. Li, "A survey of ddos attack and defense technologies in multi-access edge computing," *IEEE Internet of Things Journal*, 2024.

[30] L. N. Vandenberg and K. E. Pelch, "Systematic review methodologies and endocrine disrupting chemicals: Improving evaluations of the plastic monomer bisphenol a," *Endocrine, Metabolic & Immune Disorders-Drug Targets (Formerly Current Drug Targets-Immune, Endocrine & Metabolic Disorders)*, vol. 22, no. 7, pp. 748–764, 2022.

[31] M. R. Endsley and D. G. Jones, *Designing for situation awareness: An approach to user-centered design*. CRC press, 2025.

[32] H. E. Sari, B. Tumanggor, and D. Efron, "Improving educational outcomes through adaptive learning systems using ai," *International Transactions on Artificial Intelligence*, vol. 3, no. 1, pp. 21–31, 2024.

[33] V.-R. Bourque, Z. Schmilovich, G. Huguet, J. England, A. Okewole, C. Poulain, T. Renne, M. Jean-Louis, Z. Saci, X. Zhang *et al.*, "Integrating genomic variants and developmental milestones to predict cognitive and adaptive outcomes in autistic children," *medRxiv*, pp. 2024–07, 2024.

[34] S. Tripathi, N. Bachmann, M. Brunner, and H. Jodlbauer, "Preparedness for data-driven business model innovation: a knowledge framework for incumbent manufacturers," *Applied sciences*, vol. 14, no. 8, p. 3454, 2024.

[35] W. Zeng, X. Zhao, J. Tang, and C. Fan, "Knowledge graph alignment under scarce supervision: A general framework with active cross-view contrastive learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 9, pp. 11 692–11 705, 2023.

[36] J. D. Gates, Y. Yulianti, and G. A. Pangilinan, "Big data analytics for predictive insights in healthcare," *International Transactions on Artificial Intelligence*, vol. 3, no. 1, pp. 54–63, 2024.

[37] T. Crofts, L. Kennefick, and A. Loughnan, *The Routledge International Handbook of Criminal Responsibility*. Routledge, 2024.

[38] E. Ligia, K. Iskandar, I. K. Surajaya, M. Bayasut, O. Jayanagara, and K. Mizuno, "Cultural clash: Investigating how entrepreneurial characteristics and culture diffusion affect international interns' competency," *Aptisi Transactions on Technopreneurship (ATT)*, vol. 6, no. 2, pp. 182–198, 2024.

[39] T. Zhong, Z. Liu, Y. Pan, Y. Zhang, Y. Zhou, S. Liang, Z. Wu, Y. Lyu, P. Shu, X. Yu *et al.*, "Evaluation of openai o1: Opportunities and challenges of agi," *arXiv preprint arXiv:2409.18486*, 2024.

[40] X. Zhu, F. Lin, Z. Zhao, T. Xu, X. Zhao, Z. Yin, X. Li, and E. Chen, "Multi-behavior recommendation with personalized directed acyclic behavior graphs," *ACM Transactions on Information Systems*, vol. 43, no. 1, pp. 1–30, 2024.

[41] S. Watini, G. Davies, and N. Andersen, "Cybersecurity in learning systems: Data protection and privacy in educational information systems and digital learning environments," *International Transactions on Education Technology (ITEE)*, vol. 3, no. 1, pp. 26–35, 2024.

[42] R. Zhu, Y. Dong, J. Guo, J. He, H. Chen, R. Wang, F. Ren, A. Raben, and J. A. Martinez, "From prediction of personalized metabolic responses to foods to computational nutrition: Concepts of an emerging interdisciplinary field," *Clinical Nutrition*, 2025.

[43] N. Nuryani, A. B. Mutiara, I. M. Wiryana, D. Purnamasari, and S. N. W. Putra, "Artificial intelligence model for detecting tax evasion involving complex network schemes," *Aptisi Transactions on Technopreneurship (ATT)*, vol. 6, no. 3, pp. 339–356, 2024.

[44] M. Yan, A. P. K. Kwok, A. H. S. Chan, Y. S. Zhuang, K. Wen, and K. C. Zhang, "An empirical investigation of the impact of influencer live-streaming ads in e-commerce platforms on consumers' buying impulse," *Internet Research*, vol. 33, no. 4, pp. 1633–1663, 2023.

[45] R. Nuraeni, M. Hardini, J. Parker, and M. G. Ilham, "Swot analysis of ai-based learning recommendation systems for student engagement," *International Transactions on Artificial Intelligence*, vol. 4, no. 1, pp. 1–12, 2025.

[46] D. Tong, "Descriptor-based object-aware memory systems: A comprehensive review," *arXiv preprint arXiv:2510.27070*, 2025.

[47] J. W. Burton, E. Lopez-Lopez, S. Hechtlinger, Z. Rahwan, S. Aeschbach, M. A. Bakker, J. A. Becker, A. Berditchevskaia, J. Berger, L. Brinkmann *et al.*, "How large language models can reshape collective intelligence," *Nature human behaviour*, vol. 8, no. 9, pp. 1643–1655, 2024.

[48] Ministry of Communication and Informatics of the Republic of Indonesia, "Data-driven digital service optimization for government websites," Ministry of Communication and Informatics, Jakarta, Indonesia, Tech. Rep., 2023. [Online]. Available: https://www.kominfo.go.id

[49] A. S. Bist, "The importance of building a digital business startup in college," *Startupreneur Business Digital (SABDA Journal)*, vol. 2, no. 1, pp. 31–42, 2023.

[50] R. Hoffmann and C. Reich, "A systematic literature review on artificial intelligence and explainable artificial intelligence for visual quality assurance in manufacturing," *Electronics*, vol. 12, no. 22, p. 4572, 2023.