Jie Li, EPAM

Experimentation Everywhere and Every Day

Running A/B Testing in Corporate Environments

Imagine opening a webpage to book hotel rooms and transportation for your next holiday. As you click through pages and make selections, you might not realize that you're using just one of the tens of different variations of the webpage. You're unknowingly participating in randomized controlled experiments that contribute to companies improving your holiday planning experiences.

The fitness app you're using could be testing two notification strategies: You and the friend you're traveling with might receive different types of notifications—one emphasizing competition and another focusing on personal progress. With thousands of users contributing to data collection while using the app, the app company gains insights into which strategy triggers higher user engagement and adherence.

It's not just digital products; even your local supermarkets and department stores run tests. For instance, they might rearrange popular products at the entrance or at the back or change the direction of escalators to evaluate which variation optimizes customer shopping flows or generates more sales.

THE ROOT OF EXPERIMENTATION

The foundation of such experimentation is rooted in randomized controlled trials (RCTs) from a clinical background, which is a standard procedure for assessing whether a cause-and-effect relationship exists between medical treatments and outcomes [1]. In the basic format of RCTs, subjects—often patients—are randomly assigned to one of two groups. One group is the experimental group, where subjects receive an intervention or a new type of treatment. The other group is the control group, which receives an alternative or conventional treatment. Researchers follow up with both groups of subjects and observe whether there are differences in the outcomes.

Inheriting from RCTs, researchers in human-computer interaction also conduct controlled experiments to gain insights into human behavior, cognitive processes, physiological reactions, and interactions with new technology. Unlike RCTs, where subjects are typically patients, participants in HCI experiments are usually drawn from the general population or belong to a specific target user group for the technology under development. In contrast to RCTs, which focus on the effects of specific treatments on patients, HCI experiments often involve a broader range of independent variables (e.g., visual stimuli, devices, and interface design) and dependent variables (e.g., psychological and behavioral measurements). These measurements encompass subjective data, such as self-report questionnaires, as well as more-objective metrics like task-completion time, accuracy rates, physiological sensor data, and qualitative data, including researchers' observations and post-experiment interviews [2].

You're unknowingly participating in randomized controlled experiments.

As a specific type of controlled experiment, A/B testing is widely used in the corporate context for digital product development and user experience optimization. It typically involves well-defined A and B variations and a specific set of metrics (e.g., click-through rates, conversion rates) to measure user behavior. In today's business landscape, nearly all major companies—especially those focused on digital products and software—utilize A/B testing at scale [3]. They conduct hundreds of A/B tests on millions of users daily, covering a wide range of products, including Web platforms, mobile applications, and content arrangements for marketing campaigns. The data collected from these ongoing tests empowers companies to make data-driven decisions, rather than relying solely on the opinions of the highest-paid person.

THINGS TO BE AWARE OF WITH (AUTOMATED) A/B TESTING

Although the idea behind A/B testing is straightforward, it can become tedious when a company frequently tests a wide range of concepts and products. Some big tech companies have their in-house platforms automate the A/B testing process, including traffic allocation (i.e., the decision on how users are divided between different variations), statistical analysis, and result interpretation. This automation helps reduce manual effort and speeds up decision making (e.g., [4]). Companies that can't afford in-house solutions can leverage other commercial platforms such as VWO (https://vwo.

com/ab-testing/) and UserTesting (https://www.usertesting.com/platform/userzoom) to assist in A/B testing, enabling researchers to access real-time insights while the testing is in progress.

Although A/B testing is powerful in supporting data-driven decision making, there are drawbacks that researchers should be aware of when choosing it as a research methodology:

The Hawthorne effect. The Hawthorne effect describes how individuals alter their behavior or performance when they are aware of being observed [5]. While major tech companies automate their tests and users participate unconsciously, other companies often use online user testing platforms (e.g., UserTesting). Users participating in A/B testing through these platforms receive incentives and are often aware that their actions, though anonymous, are observed. This awareness might lead users to act favorably or even unconsciously modify their behavior.

Relatively narrow and short-term focus. The majority of online A/B testing conducted by technology companies maintains a clear goal and hypothesis. Tests are divided into atomic experiments, measuring simple user actions (e.g., the percentage proceeding to the payment page [4]). However, these tests might fail to capture whether users who didn't proceed to payment hold a positive impression of the brand and its products. Automated A/B testing requires fully functional designs, which aren't always feasible for various design phases. During the explorative design phase, inviting users to interact with low-fidelity prototypes can yield rich qualitative insights into the "why" behind their actions. A/B testing excels at testing focused sets of ideas, but isn't ideal for the exploratory design phase where concrete ideas or features haven't been defined yet.

A THOUGHT ON AI'S POTENTIAL IN A/B TESTING

As mentioned earlier, leading tech companies streamline their A/B testing processes by automating repetitive tasks through scripting.

IMAGE BY YASABII / SHUTTERSTOCK.COM

It's important to note, however, that this automation differs from the AI technologies extensively discussed today. One key distinction between automation and AI lies in their machine-learning capabilities. Automation is rule based, following predefined instructions, while AI is trained on data, capable of learning from patterns, making predictions, and even upgrading the testing scripts [6].

The data collected from these ongoing tests empowers companies to make data-driven decisions.

Recently, my colleagues and I interviewed 20 UX professionals to gain insight into their perceptions of what generative AI can and cannot do. Regarding understanding users and conducting user testing, the interviewees agreed that generative AI can help generate testing variations based on vast user behavior and preference data. It can also perform basic usability or accessibility assessments using models trained on millions of records of user data. However, it cannot replace real users participating in usability testing, nor can it validate testing results with the same level of

empathy toward users as typically possessed by UX professionals.

Essentially, AI interpolates, generating outcomes within the range of existing data. As a result, AI might excel at repetitive tasks within known data patterns. In contrast, we humans excel at extrapolating, making predictions beyond the known data range, not only using our creativity, intuition, empathy, and domain knowledge but also benefiting from multimodal grounding and the social and interactive nature of sensory input [7]. Children are good examples: They receive around four or five orders of magnitude less language data than large language models, yet vastly outperform AI capabilities [7].

AI AND A/B TESTING: POWERFUL BUT NOT EVERYTHING

As a research consultant working in a corporate setting, I often find myself trying to keep up with design sprints. Researchers need time to conduct rigorous experiments with a sufficient number of users to either explore the users' need for new features or validate design updates. However, design sprints frequently move at a rapid pace. It's common that a new design sprint has already started, while researchers have not yet concluded the testing of the design from the previous sprint. Rapid automated or AI-powered A/B testing might seem like a solution, but most design projects involve creating a brand-new service. There are often no mature, well-implemented prototypes available for A/B testing, nor a clear hypothesis about the aspects we wish to compare. A/B testing is confirmatory research, but designing new services or products typically begins with an exploratory phase.

The process of defining new features starts with a thorough understanding of the target user groups. This involves various exploratory research methods, including interviews to gather qualitative insights into user preferences, needs, and challenges; observations to uncover unexpected behaviors or interaction issues; and ethnographic studies to immerse researchers in users' natural environments and gain a deep understanding of contextual factors. Although AI has the potential to assist with some of these tedious research tasks, human researchers should lead the research efforts and validate the results. These exploratory research methods are equally essential. They help generate ideas for alternative design directions based on a comprehensive understanding of users with empathy, which can then be further explored in future A/B testing.

ENDNOTES

1. Sibbald, B. and Roland, M. Understanding controlled trials. Why are randomised controlled trials important? *BMJ: British Medical Journal 316*, 7126 (1998), 201.
2. Cairns, P., and Cox, A.L., eds. *Research Methods for Human-Computer Interaction*. Vol. 10. Cambridge Univ. Press, Cambridge, U.K., 2008.
3. Kohavi, R. and Longbotham, R. Online controlled experiments and A/B testing. In *Encyclopedia of Machine Learning and Data Mining*. Springer, Boston, 2017.
4. Glazer, A. How Booking.com A/B tests ten novenonagintillion versions of its site. Medium. Jan. 19, 2018; https://medium.com/@aaronglazer/how-booking-com-a-b-tests-ten-novenonagintillion-versions-of-its-site-25fc3a9e875b
5. Wickström, G. and Bendix, T. The "Hawthorne effect"—what did the original Hawthorne studies actually show? *Scandinavian Journal of Work, Environment & Health* (2000), 363–367.
6. Ribeiro, J., Lima, R., Eckhardt, T., and Paiva, S. Robotic process automation and artificial intelligence in industry 4.0—a literature review. *Procedia Computer Science 181* (2021), 51–58.
7. Frank, M.C. Bridging the data gap between children and large language models. *Trends in Cognitive Sciences 27*, 11 (2023), 990–992.

Jie Li is the head of research and insights at EPAM Netherlands. She has a background in industrial design engineering and her research focuses on developing evaluation metrics for immersive experiences. She is also a creative cake designer and owner of Cake Researcher, a boutique café.
→jasminejue@gmail.com

DOI: 10.1145/3636447