A Scalable Architecture for Conducting A/B Experiments in Educational Settings

Andrew Hornback
Design & Intelligence Laboratory
School of Interactive Computing
Georgia Institute of Technology
Atlanta, GA USA
ahornback6@gatech.edu

Stephen Buckley
Design & Intelligence Laboratory
School of Interactive Computing
Georgia Institute of Technology
Atlanta, GA USA
sbuckley@gatech.edu

John Kos
Design & Intelligence Laboratory
School of Interactive Computing
Georgia Institute of Technology
Atlanta, GA USA
jkos3@gatech.edu

Scott Bunin
Design & Intelligence Laboratory
School of Interactive Computing
Georgia Institute of Technology
Atlanta, GA USA
sbunin3@gatech.edu

Sungeun An
Design & Intelligence Laboratory
School of Interactive Computing
Georgia Institute of Technology
Atlanta, GA USA
sungeun.an@gatech.edu

David Joyner
Design & Intelligence Laboratory
School of Interactive Computing
Georgia Institute of Technology
Atlanta, GA USA
david.joyner@gatech.edu

Ashok Goel
Design & Intelligence Laboratory
School of Interactive Computing
Georgia Institute of Technology
Atlanta, GA USA
ashok.goel@cc.gatech.edu

ABSTRACT

A/B experiments are commonly used in research to compare the effects of changing one or more variables in two different experimental groups—a control group and a treatment group. While the benefits of using A/B experiments are widely known and accepted in education, there is less agreement on an approach to creating software infrastructure systems to assist in rapidly conducting such experiments in the field. To assist in alleviating this gap, we are creating a software infrastructure for A/B experiments that allows researchers to conduct experiments and automatically analyze their results for an education-focused ecology-based conceptual modeling platform.

CCS CONCEPTS

• **Information systems** → Data management systems; • **Software and its engineering** → Software creation and management → Designing software → Requirements analysis

KEYWORDS

A/B experiments

L@S '23, July 20–22, 2023, Copenhagen, Denmark

ACM ISBN 979-8-4007-0025-5/23/07.
https://doi.org/10.1145/3573051.3596190

ACM Reference format:

Andrew Hornback, Stephen Buckley, John Kos, Scott Bunin, Sungeun An, David Joyner, and Ashok Goel. 2023. A Scalable Architecture for Conducting A/B Experiments in Educational Settings. In *Proceedings of the Tenth ACM Conference on Learning @ Scale (L@S '23), July 20–22, 2023, Copenhagen, Denmark.* ACM, New York, NY, USA. 5 pages. https://doi.org/10.1145/3573051.3596190

1 INTRODUCTION

The A/B research methodology allows a researcher to set up a controlled experiment with two groups, typically labeled "A" and "B", one serving as a control and another as an experiment group, to test two treatments to determine which performs better. Through changing one or more well-defined characteristics in the experimental group and keeping everything else constant between the two groups, A/B experiments are used to test hypotheses about the effect of those treatments.

Specifically, A/B experiments, as a subset of normal experimental tests, benefit from and are defined by the ability to separate the users into the specified groups and collect data on their actions in real time, at any specified scale.

The A/B experimental design has been a staple in academic research for decades, and it recently has seen a surge in popularity in industry as well. As just one example, in 2012, an engineer at Microsoft working on the Bing search engine came up with an idea to change the way advertisement headlines were displayed in search results. Within hours, the new method proved its value and was able to generate insights that could increase annual revenues by more than $100 million without hindering the current user experience [1]. According to Forbes, companies that are not A/B

testing may be lagging direct competitors and companies can use results from A/B testing to make informative decisions regarding content engagement, reduce the bounce rates of users visiting their websites, and convert more customers to buyers [2].

In academia, A/B experiments are used extensively in research on learning and education and can provide benefits analogous to industry. Results of A/B experiments can help educators understand more effective ways to teach students and help students remain engaged in learning.

However, A/B experiments can require lots of manual activity; consequently, they often demand a lot of time and can be error prone. One way in which these challenges have been addressed in academia is to try and automate A/B testing. For example, Tamburrelli & Margara framed the A/B testing problem as a search-based software engineering endeavor and concluded that an automated solution for A/B experiments was possible [3]. Recently developed software tools for A/B testing in educational settings, such as UpGrade, were designed to fill voids in other off-the-shelf A/B testing systems related to issues such as the ability to assign students to a group based on criteria such as a particular school class.

The engineers who developed UpGrade specifically noted that field testing instructional improvement through A/B testing enabled highly important opportunities for learning engineers to improve learning outcomes at a faster rate than via the traditional educational research cycle [4]. Similar sentiments have been discussed by Kulkarni, who added that this "scaling through efficiency" design perspective would allow instructors to assist a larger number of students [5].

As part of the scarce research into this topic, Chudzicki has shown the possible benefits of A/B testing in educational environments by running experiments using Edx's on platform A/B testing tools [6]. The Edx A/B testing platform is limited to course content, which limits the variety of possible applications. While not all researchers and organizations have the resources that large technology companies such as Microsoft possess, they all can benefit directly from A/B experiments in the same manner. To do so, a software architecture that can easily enable A/B experiments has the potential to serve as a mechanism to reduce gaps related to constraints on resources. However, it is not enough to build software that can support learning, in the case of academia, or consumer analysis in the case of industry. By enabling A/B experiments to be easily conducted, such systems must have built-in components to make analytical insights easily understandable and obtainable by users.

Despite the importance of these goals and the success of A/B experiments, there is only modest research on a scalable approach to developing software architectures for conducting A/B experiments. There exists a need for such an approach, as it would assist researchers and organizations with creating A/B experiment software quickly. In the paper, "Improving Library User Experience with A/B Testing: Principles and Process" Scott W.H. Young [7] outlines a step-by-step framework for conducting A/B testing, and it includes the following steps:

1. Define a research question.
2. Refine the question with user interviews.
3. Formulate a hypothesis, identify appropriate tools, and define test metrics.
4. Set up and run an experiment.
5. Collect data and analyze results.
6. Share results and make decisions.

We posit that any automated A/B testing software should enable researchers to conduct all these steps. As such, the software we developed is designed to minimize the burden on the researcher by automatically handling steps four and five, allowing the researcher to focus on the more foundational steps of one and two. Additionally, steps three and six are partially covered by our design by giving additional test metrics and allowing the researcher to share results.

2 VERA

VERA is a virtual laboratory for learning about the scientific way of thinking. Since 2007, VERA has been used for a variety of learners in multiple settings. Teacher-guided middle school science courses, laboratory sections of college undergraduate biology courses, REU summer schools, summer internship programs at museums, adult learners in online programs, citizen scientists seeking to make sense of environmental data, and globally distributed learners of unknown demographics engaged in self-directed learning have all used VERA.

In addition, Smithsonian Institute's Encyclopedia of Life (EOL) website (http://www.eol.org) provides direct access to VERA to millions of visitors each year. Lifelong learning is one of the longstanding goals of education [8, 9] and is expounded by VERA through its free and open access to all learners.

Learning about the scientific way of thinking is another long-standing goal of education [10, 11]. VERA is an inquiry-based modeling environment designed to enable learners to explore ecological and other complex systems by performing "what-if" experiments to either explain the behavior of an existing system or attempt to predict the outcome of structural changes to one. Learners use conceptual models and run agent-based simulations of these models to conduct these experiments while engaging in the scientific way of thinking.

3 DEVELOPING A/B EXPERIMENT ARCHITECTURE

To expand VERA from an online laboratory to a tool capable of enabling researchers and educators to conduct rapid A/B experiments with different groups of learners in different learning contexts, we designed and implemented software to automatically conduct experiments and run corresponding analysis.

The expanded version of VERA features a Researcher mode that encompasses infrastructure that collectively meets the requirements without sacrificing the established user experience that has been part of VERA historically. Model construction, agent-based simulation, hypothesis testing, and all other aspects of VERA

did not require direct modifications to implement the new software features designed for A/B experiments.

To conduct an A/B experiment using the new infrastructure in VERA, researchers begin by creating an experiment using VERA's web interface. Instructions are provided and a researcher provides a name and description of the experiment that will be used to hierarchically organize the given experiment in conjunction with other experiments the researcher conducts.

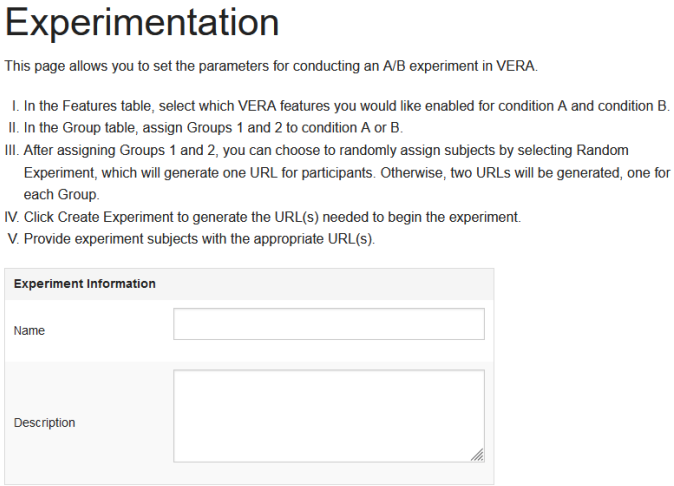

Figure 1: When creating an experiment, researchers are provided instructions as shown, and they begin by inputting the name and description of the experiment.

While creating an experiment, the researcher can enable and disable certain features for the two experiment groups. These features include advanced parameters, cloning, exemplar models, lookup EOL, and simulation. Advanced parameters are values associated with model components and include photosynthesis rate, assimilation efficiency, move velocity, respiratory rate, move direction, and carbon biomass. Cloning involves taking a pre-existing model and making a copy to initialize a new model from. Exemplar models are template models available in VERA that are often cloned. Lookup EOL allows users to retrieve parameter values for a biotic component from the Encyclopedia of Life as opposed to initializing the parameters on their own. Finally, simulation allows users to run the models constructed.

Paired with the ability to randomly or manually assign experiment participants to an experiment, this part of the software infrastructure design helps to ascertain the first principle proposed as a design requirement, retaining the typical properties of A/B experiments.

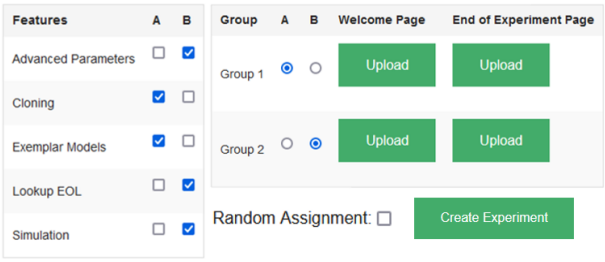

Figure 2: After filling out the fields shown in Figure 1, researchers can enable or disable features of VERA to set up their experiment. They first assign on/off features to Condition A and Condition B (as shown on the left), then they assign one of the conditions to each of the two experiment groups, Group 1 and Group 2 (as shown on the right). Researchers can also upload welcome and end pages for the experiment and choose whether subjects should be randomly assigned to groups.

4 CONDUCTING AND ANALYZING AN A/B EXPERIMENT

After a researcher creates an experiment and sets the conditions for both the control and treatment group, one or two URLs will be generated depending on whether the experiment was randomly or manually assigned. The researcher provides experiment participants with the URL(s) which lead to modified versions of VERA based on the settings chosen by the researcher in the experiment design (for example, if the researcher chose to remove Simulation and no other features for Group A, the URL provided to Group A would lead to a version of VERA without this feature).

Group	Link
1	https://vera.cc.gatech.edu/researcher/join-experiment?group=143
2	https://vera.cc.gatech.edu/researcher/join-experiment?group=144

Figure 3: Once researchers create an experiment, they are provided with URLs that lead to modified versions of VERA. Researchers then provide these URLs to participants in their experiment groups to begin conducting an experiment.

While each experiment will encompass the unique design of the researcher's desired hypothesis to be tested, the data capture and management system of VERA is experiment-agnostic and provides researchers with tools needed to complete their experiment accordingly. Problem-solving activities and outcomes are captured in VERA's backend database. All information related to experiment features, such as Simulation, are automatically stored and paired with a set of analytics metrics that assist the researcher in understanding experiment participant behavior that shapes the outcome of the experiment.

For example, models constructed in VERA differ in the number of components and the uniqueness of the components used to solve

a problem. A learner may place two biotic components and a relationship, such as consumes, between the biotics to model a system. Another learner may model the same system using two biotics and an abiotic, and a differing set of relationships among the three components. To help researchers understand the difference in such behaviors, two custom metrics are available to researchers for each user at the end of an experiment: model complexity (the sum of components used to construct a model) and model variety (the sum of unique components used to construct a model). These custom metrics are paired with traditional descriptive statistics such as the total number of learners in an experiment and automatically provided to researchers in numerical and visual form to exemplify the enabling automated analysis principle.

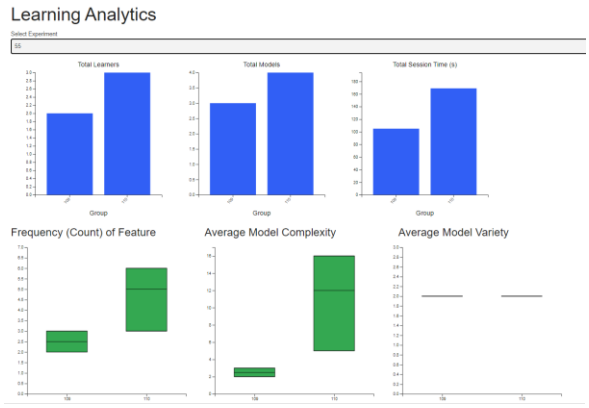

Figure 4: Researchers are provided standard and VERA specific analytic metrics automatically at the end of each experiment.

Additionally, further statistics are tracked and modeled for each user, although they remain anonymized. The two user-specific metrics are Markov Chains and Activity Sequences, both of which detail the flow of user actions through the system. These models were included because of their previous use in learning research on the VERA platform [12]. In previous research, this primarily took the form of the Activity Sequences; however, the Markov Chain illustrates the same data as a probability that a user will go from one action to the next instead of directly enumerating each user action.

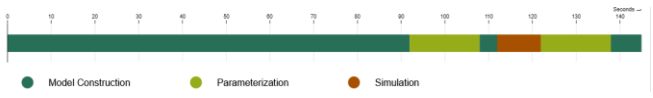

Figure 5: All actions are categorized into one of three categories: model construction, parameterization, and simulation. Then, the shown Activity Sequences demonstrate the order (and amount of time in seconds) in which a user spent completing actions within these categories while working on a model.

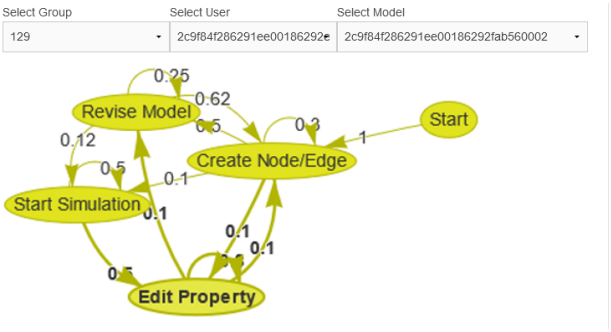

Figure 6: As shown, Markov Chain visualizations in VERA demonstrate common sequences of actions a user took while working on a model. Each node is an action, with an example being starting a simulation, and each edge (and corresponding weight) represents the probability of completing one action after completing an initial one. Within this context, Markov Chain Visualizations particularly serve to illustrate common loops in user actions.

5 CONCLUSIONS AND FUTURE WORK

A/B experiments provide an efficient way for researchers to understand the difference in one or more variables using a method with robust historic support. We have developed a system for researchers that allows them to design experiments using VERA and conduct them with minimal setup time and a software infrastructure that assists with analysis. We believe our approach can be expanded to other systems as well.

The new Researcher mode is part of a continuous effort to expand VERA in terms of both functionality [13] and accessibility to educators. We already have deployed and validated the software infrastructure in a graduate class at our university. We have formed relationships with a technical college where we plan to help deploy the infrastructure in courses related to natural resource management in the upcoming Summer and Fall 2023 semesters. The results of the experiments and the feedback provided by the course instructors will allow us to evaluate the validity of our architecture and better evaluate its generalizability, while also providing us with information that can be used to improve the system.

Using the feedback gathered, we aspire to expand the research engagement of VERA to additional educational settings that enable large-scale insights into learning behaviors. Ultimately, we are working toward providing a domain-agnostic framework for others to conduct rapid A/B experiments at scale.

6 ACKNOWLEDGMENTS

This research is supported by the National Science Foundation under Cooperative Agreement DRL-2112532 with the National AI Institute for Adult Learning and Online Education (aialoe.org). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] R. Kohavi and S. Thomke, "The Surprising Power of Online Experiments," Harvard Business Review, 16-Sep-2020. [Online]. Available: https://hbr.org/2017/09/the-surprising-power-of-online-experiments. [Accessed: 26-Sep-2022].

[2] J. Simpson, "A/B Testing: The Benefits And How To Use It Efficiently," Forbes, 12-Mar-2020. [Online]. Available: https://www.forbes.com/sites/forbesagencycouncil/2020/03/12/ab-testing-the-benefits-and-how-to-use-it-efficiently/?sh=49de0e7786d4. [Accessed: 22-Sep-2022].

[3] Tamburrelli, Giordano & Margara, Alessandro. (2014). Towards Automated A/B Testing. 10.1007/978-3-319-09940-8_13.

[4] Ritter, Steven & Murphy, April & Fancsali, Stephen & Fitkariwala, Vivek & Patel, Nirmal & Lomas, Derek. (2020). UpGrade: An Open Source Tool to Support A/B Testing in Educational Software.

[5] C. Kulkarni, "Design Perspectives of Learning at Scale: Scaling Efficiency and Empowerment" L@S '19: Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale, June 2019 Article No.: 18 Pages 1–11 https://doi.org/10.1145/3330430.3333620

[6] Chudzicki, Christopher, David E. Pritchard, and Zhongzhou Chen. "Learning experiments using AB testing at scale." Proceedings of the Second (2015) ACM Conference on Learning@ Scale. 2015.

[7] N Young, S. W. (2014). Improving library user experience with a/B testing: Principles and process. Weave: Journal of Library User Experience, 1(1). https://doi.org/10.3998/weave.12535642.0001.101

[8] Joyner, D., and Goel, A. 2015. Improving Inquiry-Driven Modeling in Science Education Through Interaction with Intelligent Tutoring Agents. In Procs. 20th ACM Conference on Intelligent User Interfaces, 5-16. New York: ACM

[9] Field, J. (2000) Lifelong learning and the new educational order. Trentham Books, UK.

[10] Kuhn, D., Amsel, E., & O'Laughlin, M. (1988). The development of scientific thinking skills. Orlando, FL: Academic Press.

[11] Lehrer, R., & Schauble, L. (2015). The development of scientific thinking. In L. Liben & U. Mueller (Vol. eds.) & R.Lerner (Series ed.), Handbook of child psychology and developmental science, Vol. 2: Cognitive process. (7th Edition). Hoboken, NJ: Wiley.

[12] An, S., Rugaber, S., Hammock, J., Goel, A.K. (2022). Understanding Self-Directed Learning with Sequential Pattern Mining. In: Rodrigo, M.M., Matsuda, N., Cristea, A.I., Dimitrova, V. (eds) Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners' and Doctoral Consortium. AIED 2022. Lecture Notes in Computer Science, vol 13356. Springer, Cham. https://doi.org/10.1007/978-3-031-11647-6_102

[13] Scott Bunin, Willventchy Celestin, Andrew Hornback, and Spencer Rugaber. 2022. Incorporating Habitats in Conceptual Models and Agent-Based Simulations: Expanding the Virtual Ecological Research Assistant (VERA). In Proceedings of the Ninth ACM Conference on Learning @ Scale (L@S '22). Association for Computing Machinery, New York, NY, USA, 472–474. https://doi.org/10.1145/3491140.3528261