Building an Infrastructure for A/B Experiments at Scale: The Challenges, Opportunities, and Lessons for the Learning Analytics Community

Sreecharan Sankaranarayanan, Chellie Harrison, Sarita Kumari, Kim Larson, Robert Lemiesz, Nicolas Mesa, Ryan Mitts, Ishita Verma, Dawn Zimmaro
Talent Development, Amazon.com, Inc.
{sreeis, chelliec, saritku, kilarson, lemiesz, nicfern, rmitts, eshaverm, dzimmaro}@amazon.com

ABSTRACT: A/B testing at scale provides opportunities for learning analytics researchers to learn from large sample sizes. Deploying and running live intervention experiments with such large samples, however, raises infrastructural challenges. This paper discusses some of those challenges, and reports on two possible implementations that address those challenges in a workforce learning context at a large technology company. The pros and cons of the alternatives are discussed with the help of a specific use-case, an A/B test comparing personalized feedback on open-ended responses to static feedback. Along the way, the paper discusses idiosyncrasies that have to be kept in mind while conducting and evaluating learning experiments in the industry.

Keywords: A/B Testing, Workforce Learning, Learning at Scale, Experimentation at Scale

1 MOTIVATION

A/B testing (terminology emerging from user experience research) is a "between-subjects" experimental method where an individual is assigned to one of two (or more) experimental conditions (often control, and one or more treatment conditions). Provided the baseline characteristics of these groups are identical across conditions, any observed effect can be attributed to the treatment. While a random assignment of individuals to conditions is often used to create groups with similar baseline characteristics, a small sample size reduces the statistical power of generalizing the findings to a broader population outside of the study (Faber & Fonseca, 2014). This is the promise of A/B testing at scale. A large sample size increases the likelihood of observed effects generalizing to the population.

The promise of A/B testing at scale is, of course, complemented by several challenges. First is the challenge of finding an appropriate use-case – one that measurably and quickly advances business and/or learning goals while also satisfying the dual methodological requirements of limiting possible harm or inconvenience from a treatment, as well as preventing a beneficial treatment from being withheld for too long. Second is the technological challenge of developing, testing, deploying, and maintaining an infrastructure that enables A/B testing experiments to be conducted at scale. Last is the technological challenge of collecting the data, quickly analyzing it, and, preferably automatically, declaring a "winning" condition either to limit harm or inconvenience, or to multiply the benefit.

Both large samples and the associated infrastructure for A/B testing at scale have previously been relegated primarily to large-scale collaborative initiatives such as the Pittsburgh Science of Learning

DataShop (Koedinger et al., 2010), the Super Experiment Framework under the Next Generation Learning Challenge (Stamper et al., 2012), E-TRIALS from ASSISTments (Krichevsky et al., 2020), and MOOCs (Reich, 2015), although noteworthy examples from single research groups do exist (Mostow et al., 2003). This paper describes two kinds of implementations built to address the challenges described above. The first uses custom-built or equivalent open-source technologies, and the second uses Amazon Web Services (AWS) technologies. The pros and cons of the two implementations are discussed with the help of an example use-case comparing personalized, AI-generated feedback on open-ended responses to static, hard-coded feedback. The discussion should provide research groups with the information needed to build scalable A/B testing infrastructures of their own.

2 THE USE-CASE

The industry context provides us access to a large sample but requires additional considerations. An ideal use-case will be one that produces actionable results in the short-term. That is not to say that long time-horizons are not possible, but that they need to be coupled with delivering results in the short-term. One such use-case was identified in an eLearning course for learning designers on writing learning outcomes. An open-ended question asked learners to rephrase and improve on a sub-par learning outcome. The control condition provided learners with an exemplar that they can, on their own, compare against their response. Building on prior work (Zhao et al., 2021), the treatment condition provided learners with personalized feedback on their response using an algorithm that sits atop a pretrained language model. Given a set of learner answers and exemplar answers, the algorithm identifies key phrases present in the model answers that are missing from the learner answer, allowing those missing key phrases to be delivered as personalized feedback to the learner. The model backend additionally determines a score for the learner answer based on this comparison between their answer and the model answers. This score is not displayed to the learner and is only used for hypothesis testing. In either condition, learners can submit responses as many times as they want, although only the treatment condition would provide personalized feedback each time. Four hypotheses were tested across the conditions. First, we hypothesized that learners who received personalized feedback would be more likely to edit their responses and resubmit. This is fairly trivial; however, it enables us to test if the additional time spent revising the responses was worth it for learning. Second, as a result of the personalized feedback, the quality of the terminal submission, as measured by the model-generated score, ought to be higher in the treatment. Third, the combination of the personalized feedback, and an increased number of attempts to refine their submission based on it, should lead to improved learning in the treatment. Finally, the model-generated score should be higher in the treatment condition for equivalent attempt numbers (second attempts compared across conditions, for example). The final hypothesis allows us to understand the mechanism of learning i.e., whether learning was due to personalized feedback or due to editing and refining one's responses. Lacking a proximal measure of learning such as a pre- and post-test, we used performance on the set of assessments tied to the learning outcome associated with the open-ended response question as our (distal) measure of learning. Since it is a high bar to expect a single question to produce a significant impact on the learning outcome as a whole, we reserve the last two hypothesis as exploratory. The results of the first two hypotheses, therefore, determines the "winner" of the A/B test, with the experimental condition being declared as such only if both of the first two hypotheses are satisfied.

3 THE INFRASTRUCTURE

Open-source frameworks such as Upgrade allow augmenting existing educational applications with A/B testing capabilities through Learning Tools Interoperability (LTI). Nevertheless, they are unsuited for use in corporate contexts for various reasons – scalability, security, and lack of customizability/extensibility being chief among them. As our first solution, therefore, we built a custom solution that is similar to Upgrade. Two copies of the learning experience, one with the desired treatment, are made, and the system redirects learners to one or the other based on the rules provided through an authoring interface. Note that the authoring interface needs to be implemented in addition to the redirection logic. The authoring interface can specify the percentage of learners in each condition, defined start and end conditions, and the post-experiment actions among other parameters. The redirection logic is not only a one-time redirection of a learner to an experimental condition but has to persist that assignment across learner sessions through drops in connection, log-outs, and breaks. Data from each condition must then be collected, and correctly and consistently mapped to the learners assigned to those conditions. If data-processing rules are set, the evaluation metrics can be automatically calculated to determine the outcome of the experiment and the "winning" condition that subsequent learners will be assigned to. Otherwise, duration or number of learners can also be used as terminating conditions. In our case, the number of learners (200-400) is used as the termination condition while we wait for the automatic evaluation to be implemented. The evaluation is conducted offline, and learners are reverted to the control condition in the meantime.

Our second implementation leverages Amazon Cloudwatch Evidently. Features for both the control and treatment are implemented, but hidden behind 'if' statements on the front-end interface. At runtime, the application code queries a remote service. The service decides the percentage of users who are exposed to the new feature, i.e., the treatment, and returns information about whether a specific specific user is in the treatment or control condition allowing the appropriate front-end components to be rendered. Evidently additionally allows statistical analysis to be performed at run-time and decisions made based on the results of those statistical analyses.

Each approach has its pros and cons. The former learning experience-level redirection approach allows more complex A/B comparisons such as different sequences of instruction and assessment, while the latter real-time rendering approach is ideally scoped to the level of a page, or even one learning object on the page. On the other hand, the experience-level redirection approach causes duplication of assets resulting in an increase in the burden of reporting on metrics such as completion, and an increased likelihood of learners chancing directly upon the URL for the control or treatment learning experiences in case the owner omits updating the link in all the places it is referenced (email blasts, wiki links, course catalogs etc.). The real-time rendering implementation does not require a redistribution of links and allows for a smoother post-experiment transition since it is all handled within the same learning experience. The determination for which infrastructure to use will depend on the requirements of the experiment. In our case, since the experiment focused on rendering a single assessment within the learning experience, the real-time rendering approach was the best suited.

Another element of this experiment is the machine learning model consuming learner answers and providing real-time personalized feedback. Not only does this involve the backend infrastructure

necessary to consume learner answers, return the model output, and fashion that into a response to learners, it also involves the frontend infrastructure elements of rendering the personalized feedback, refreshing the interface to allow learners to respond again, and re-rendering feedback. All of this has to, of course, happen in real-time to allow for a reasonable experience for learners. The model inference code is easy to deploy, packaged along with all the requirements files into a Docker container that is then deployed using Amazon SageMaker. The front-end component development is much more effortful.

4 ETHICS OF A/B

The A/B infrastructure has interesting implications for the ethics of A/B experiments, which we wanted to make explicit. A well-developed infrastructure will allow for quickly resolving in favor of the experiment or treatment conditions, thus preventing harm or inconvenience, or making the beneficial treatment available. Of course, this is ultimately down to the experimental design and the experimenter. Nevertheless, the needed ease of doing so has implications for the design of the A/B infrastructure, as outlined in earlier sections.

5 CONCLUSION

We present two possible infrastructures for A/B testing learning experiences at scale, discuss their pros and cons, and demonstrate making a choice between them with a use-case. We expect research groups to be able to use this information to build scalable A/B testing infrastructures of their own.

REFERENCES

Faber, J., & Fonseca, L.M. (2014). How sample size influences research outcomes. *Dental Press Journal of Orthodontics, 19*, 27 - 29.

Glaser, A. (2021, July 30). *Amazon now employs almost 1 million people in the U.S. - or 1 in every 169 workers*. NBCNews.com. Retrieved December 8, 2022, from https://www.nbcnews.com/business/business-news/amazon-now-employs-almost-1-million-people-u-s-or-n1275539

Koedinger, K. R., Baker, R. S., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A data repository for the EDM community: The PSLC DataShop. *Handbook of educational data mining*, *43*, 43-56.

Krichevsky, N., Spinelli, K., Heffernan, N., Ostrow, K., & Emberling, M. R. (2020). *E-TRIALS* (Doctoral dissertation, Doctoral dissertation, Worcester Polytechnic Institute).

Mostow, J., Beck, J. E., & Valeri, J. (2003). Can automated emotional scaffolding affect student persistence? A Baseline Experiment. In *Proceedings of the Workshop on" Assessing and Adapting to User Attitudes and Affect: Why, When and How?" at the 9th International Conference on User Modeling (UM'03)* (pp. 61-64).

Reich, J. (2015). Rebooting MOOC research. *Science*, *347*(6217), 34-35.

Stamper, J. C., Lomas, D., Ching, D., Ritter, S., Koedinger, K. R., & Steinhart, J. (2012). The Rise of the Super Experiment. *International Educational Data Mining Society*.

Zhao, J., Larson, K., Xu, W., Gattani, N., & Thille, C. (2021). Targeted feedback generation for constructed-response questions. *AAAI-2021 Workshop on AI Education*.