A Progress Report & Roadmap for A/B Testing at Scale with UpGrade

Leave Authors Anonymous
for Submission
City, Country
e-mail address

Leave Authors Anonymous
for Submission
City, Country
e-mail address

Leave Authors Anonymous
for Submission
City, Country
e-mail address

ABSTRACT

This paper serves as a progress report and roadmap for UpGrade, an open source A/B testing platform for digital experimentation in educational technologies. The UpGrade roadmap focuses on reducing barriers to user onboarding and rapidly providing value to the learning engineering community by leveraging existing technical standards. We briefly describe UpGrade, lay out our roadmap, and discuss several planned UpGrade features. We intend for this paper to spark discussion and feedback from the learning engineering community as to course corrections for the roadmap, standards that can be fruitfully leveraged, and features that would enable more educational technologies to rigorously, experimentally test product improvements and evaluate educational effectiveness.

Author Keywords

A/B testing, Educational technology, Digital experimentation

INTRODUCTION

UpGrade is a free, open-source A/B testing platform that enables large-scale field experiments with educational software applications. It functions as a web-based service for designing and handling logistics for randomized trials, interacting with but operating separately from the software that delivers educational content and any additional application service(s) (such as an LMS) that manage the software's user data. To use the platform, researchers or educational technology software developers host an instance of the UpGrade server, either locally or via a cloud-based service. UpGrade then integrates with the desired educational platform using a client library. Once integrated, the UpGrade user interface provides researchers or experimenters with the ability to set appropriate parameters and provides user analytics from the integrated educational app for metrics that are set for "monitoring" in the integration process (Ritter et al, 2020a).

CHI 2020, April 25–30, 2020, Honolulu, HI, USA.
 ACM ISBN 978-1-4503-6708-0/20/04...$15.00.
DOI: https://doi.org/10.1145/3313831.XXXXXXX
*update the above block and DOI per your rightsreview confirmation (provided after acceptance)

Figure 1. Monitoring enrollment in an UpGrade experiment, showing the number of students assigned to each condition and enrollment over time by condition.

A key feature of UpGrade is that it provides logic to handle group random assignment at scale. Individual randomization is common across A/B testing platforms, but when conducting experiments in educational contexts, the ability to assign students to a condition according to group membership (such as class, teacher, or school) is both highly desirable and necessary to provide consistent learning experiences to students within such groups. This functionality is particularly valuable for experiments that test educational experiences that substantially differ, perhaps by learning model or visual design, as it limits the burden on teachers to keep track of which students receive which approach. Such functionality also reduces possible perceptions of inequity or unfairness by students themselves if they observe different software experiences among one another (Ritter et al, 2020b).

One of the challenges of handling group randomization at scale is the management of real-world classroom scenarios that may violate group membership or consistency. UpGrade's configurable parameters allow researchers to set "consistency rules" for handling such scenarios, which are relatively common within educational institutions but not well-handled by off-the-shelf A/B testing platforms. The ability to adjust such consistency rules is of particular advantage in adaptive or self-paced educational software, when students may not reach educational content at the same time. Similar concerns about whether and how to enforce within-class consistency with respect to condition assignment can apply if a student is enrolled in multiple classes. Students may also change group membership, and rules can be set to decide whether, if class A receives

learning experience A and class B receives learning experience B, which learning experience should we provide the student who transfers from class A to class B? UpGrade recognizes the need to address these and related situations when designing A/B tests for groups; further details of group assignment considerations are described in Ritter et al, 2020b. While UpGrade fully supports designs for non-school and individual contexts, the platform is set apart from alternatives by its prioritization of the capacity to conduct group random assignment experiments in real-world classroom settings at scale.

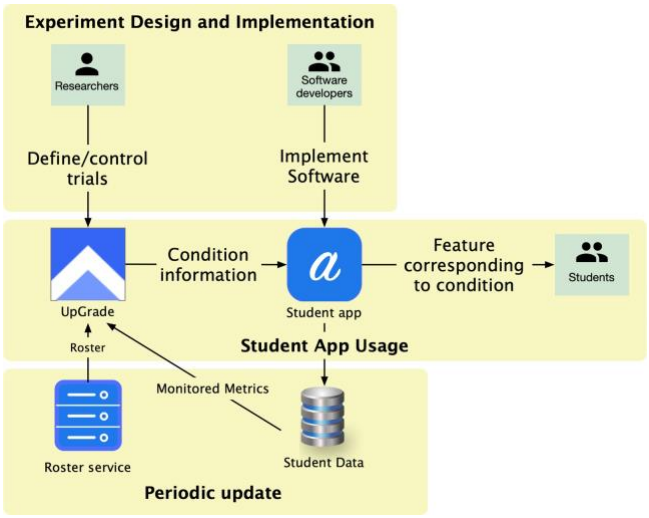

Figure 2. Overview of UpGrade Architecture

Leveraging technical standards to reduce implementation barriers

Designing an experiment and setting rules for managing condition-assignment behavior are important steps in the goal of running A/B tests at scale, but there are practical details to take into consideration during the implementation stage. For experiments involving group assignment, we need a way to provide roster information to UpGrade. To monitor metrics within the UpGrade UI, we need to share data between UpGrade and the educational app to which it connects. For partners at research institutions whose designs may involve comparing courses or products from multiple sources, we require a way for the UpGrade server to behave as if the institution's Learning Management System (LMS) were the educational application, with appropriate processes in place to deliver assigned conditions within the LMS platform. By incorporating technical standards in UpGrade's implementation model we can provide educational software developers and researchers pre-written software that performs these functions, reducing the complexity of implementation and limiting the amount of custom code to write. Standards-based adherence to established practices should also support reliability and encourage broader usage.

Data capture and sharing

Within UpGrade, an experimenter can request metrics, such as error count, time to complete a learning module, or others appropriate to the experiment, then specify the desired simple statistic to compute (e.g. mean, median, maximum), as well as logic such as whether metrics should be calculated by student, grouped, and how repeated measures (for instance, a student returning to a learning module in review mode) should be handled. To support this functionality,, we defined a proprietary format to share data between the educational application and UpGrade. The format is quite general, but, since each application stores the data in its own way, each application using UpGrade needs to build a custom pipeline to convert their data into UpGrade's format. This pipeline code must take the data structure generated from these specifications, communicate with the learning app, and return the appropriate data to UpGrade.

For applications using a standard data format like xAPI, we can provide code that is configurable to perform this function without additional coding. Experience API (xAPI) is a leading choice because of its position as a standard specific to e-learning, and because it can capture data from many different learning technologies, such as mobile apps, games, offline learning, virtual reality, and simulations, among others (the IMS Project's Caliper might also be a target of this work). The API captures information about a person or group's learning activities and encodes this in a format that allows for easy movement of data between an application and UpGrade.

Using roster data for group assignment

As noted earlier, UpGrade's most unique feature may be its capacity to facilitate group random assignment at scale. Even though UpGrade itself does not handle or store personally-identifiable information, it must have a way of accessing student group membership in order to effectively assign conditions to groups. This can be accomplished via communication with a student roster that stores relevant data (e.g., class, teacher, school, or district). There are two models in which UpGrade can acquire roster information. One approach considers the educational application connecting to UpGrade as the de facto roster service. This is how UpGrade is implemented with Carnegie Learning's MATHia software. If the app itself stores student group information, this approach can work well. A potential downside is that UpGrade's view of roster information can only be updated when a student logs in to the app. Scenarios such as infrequent student logins combined with class transfers could lead to the roster being populated with obsolete data, resulting in undesirable consequences such as other students in the class being unnecessarily excluded from the experiment.

A second approach is for UpGrade to communicate with an external rostering service that stores and updates student

information such as class and other types of group memberships. Because such services are not reliant on student behavior to trigger data updates, the possibility of information obsolescence impacting group condition assignment is effectively eliminated, since the roster can update UpGrade as often as required, according to pre-set configuration parameters. OneRoster is one of several models for handling roster data structures, and is used by many educational technology products and LMSes. Adapting UpGrade to support a standard data model such as OneRoster should streamline this component of implementation, particularly for partners interested in group random assignment.

Using LTI to enable randomized experiments at the application level

Researchers who wish to conduct randomized comparisons of learning experiences are not always interested in user experiences within a single application; rather the goal may be to evaluate learning outcomes from *different* applications or sources. A proposed solution is to allow users to run UpGrade experiments with an institution's LMS standing in for the educational application. In this model, an LTI link inside the LMS's Common Cartridge could pass user information into UpGrade via an authentication service. After collecting this data, UpGrade can then determine condition assignment, then refer back to the service to direct students to appropriate condition URLs within the LMS. See Figure 3 for a diagram of how we plan to integrate these processes.

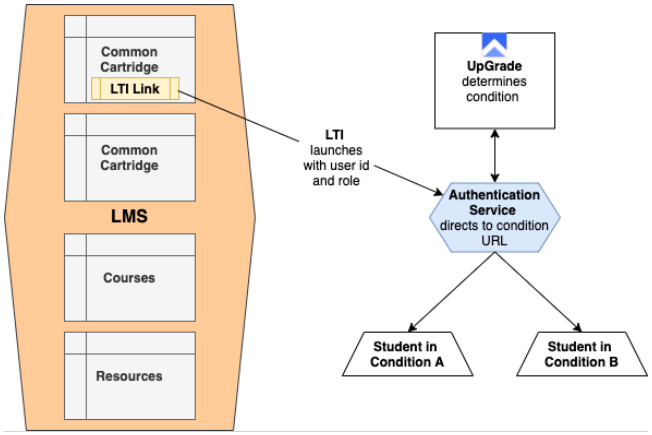

Figure 3. Diagram demonstrating how UpGrade can interact with an LMS in place of the educational app. LTI communicattes with UpGrade to assign conditions as URLs within the LMS.

Feature Development

Beyond adopting technical standards to support implementation efforts, our roadmap for UpGrade includes three new features/enhancements: an interactive application for improved outreach and demonstration of UpGrade's capabilities, an user-friendly interface for experimenters to set include/exclude rules for segmenting users, and a robust statistical approach to acquiring and balancing student-level demographics from randomized trials.

Interactive web application

There is no single prototype that defines a typical UpGrade user-experimenter. UpGrade A/B testing can be carried out by educational technology companies, academic researchers, school districts and even teachers. We have previously focused on software developers in the ed tech space, but one way to broaden applicability to the wider community is to explicitly highlight use cases relevant to varied audiences and their goals. As part of expanding the UpGrade website, we are building an interactive, web-based demo whereby interested visitors to the site can log on to an instance of UpGrade and explore setting up experiments with different types of learning experiences (for instance, content and features changes) and use cases from worked examples. The demo will provide an end-to-end snapshot from initial design to observing how participant assignment and monitored metrics are populated as they would be in a typical application.

User segmentation

Targeting population segments that share common attributes is a standard approach to understanding user behavior in web marketing. The same tactics can be used to understand specific types of learner behavior in educational product development. Segmentation rules can also be applied when researchers have agreements or IRB approval only with specific schools or districts. UpGrade developers are creating a segmentation feature that will enable experimenters to easily create include/exclude rules for populations based on predefined characteristics (for instance, region, grade level, district, school-level demographics) to narrow the focus of experiments to desired targets. This feature is one component of a two-part objective to enable more robust demographics acquisition in UpGrade experiments. We discuss the second component in the next section.

Student demographics in large-scale A/B testing

Education researchers have long been interested in assessing learning outcomes from subgroups such as BIPOC students, students experiencing poverty, groups with special learning needs, and other hard-to-reach populations. In general, student-level demographic and school record data is protected under FERPA, and its usage is prohibited for general research purposes without IRB and/or parental approval. School-level demographic information, however, is publically available through sources like NCES. The primary drawback here is that statistical power is reduced by using school-level, rather than individual-level subgroup characteristics when analyzing learning outcomes. The UpGrade platform

neither collects nor stores student data (acquiring any such demographic information from the roster service of the educational application to which it connects), so designing A/B tests to address these groups can present practical and methodological challenges, as student demographics are used in different ways for different reasons. In this section we discuss two situations in which student-level demographics are currently used in A/B testing, and propose a third approach that can address the unique challenge of acquiring balanced subgroup data from sequential random assignment.

In some cases, student-level achievement, race, disability status, and other PII are part of the purpose of the A/B intervention and thus permitted through FERPA under these circumstances. For instance, if an intelligent tutoring program has an intervention that supports students with dyslexia, disclosure to the software provider about which students have learning disabilities may be permitted. In this case, the demographic information is part of the learning experience intervention itself, and is necessary for it to function as intended.

Another way student-level demographic information is used is when it is part of a planned analysis. Schools who maintain explicit data-sharing agreements with companies that produce educational software used by the school can provide their student data to the company under such an agreement. This data can be used as part of other analyses conducted on the results of A/B educational interventions.

When student-level demographics are used in random assignment, the primary challenge is to achieve a well-balanced design (that is, approximately equal participants assigned per condition) among subgroup characteristics of interest. Blocking by school to balance subgroups can be effective when group randomization occurs at the beginning of the study and assignment is at the teacher or class level, but sequential random sampling--when groups are randomized between conditions one at a time--presents a more difficult problem. This is frequently the case with educational software, when schools may begin using the software at any point during the year, or when software is adaptive and students reach nodes in an educational sequence asynchronously. Here we propose to integrate "biased-coin" designs (Efron, 1971; Antognini & Zagoraiou, 2011) with UpGrade to promote balance by adjusting treatment probabilities at different stages of randomization. This novel approach, originally used in biomedical research, will be adapted for use in educational contexts as stratified, group-randomized trials with covariates of relevance to education. A similar challenge will be to adapt treatment effect estimation methods to biased-coin designs (Ma, et al., 2019).

CONCLUSION

This paper describes a roadmap for UpGrade, an open-source platform for large-scale A/B testing in educational applications. We propose to incorporate a range of technical standards, features, and enhancements to reduce implementation barriers and benefit the learning engineering community as a platform that can enable rigorous, large-scale testing of learning experiences and their effectiveness. As UpGrade gains traction and an increased user base, we expect future roadmaps and feature development to be a collaborative process among user-developers and other stakeholders. UpGrade source code is available on GitHub. If you are interested in using or contributing to UpGrade, visit www.upgradeplatform.org or email upgradeplatform@carnegielearning.com.

ACKNOWLEDGMENTS

This work was supported by grants from the Bill & Melinda Gates Foundation and Schmidt Futures.

REFERENCES

[1] Antognini, A., & Zagoraiou, M. (2011). The covariate-adaptive biased coin design for balancing clinical trials in the presence of prognostic factors. *Biometrika*, 98(3), 519-535.

[2] Efron, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika*, 58, 403.

[3] Ritter, S., Murphy, A., Fancsali, S., Fitkariwala, V., Patel, N., & Lomas, J. D. (2020). UpGrade: An Open Source Tool to Support A/B Testing in Educational Software. In *L@S Workshop on A/B Testing at Scale*.

[4] Ritter, S., Murphy, A., Fancsali, S. (2020). Managing Group Random Assignment in UpGrade. In *L@S Workshop on A/B Testing at Scale*.

[5] Ma, W., Qin, Y., Li, Y., & Hu, F. (2020). Statistical inference for covariate-adaptive randomization procedures. *Journal of the American Statistical Association*, *115*(531), 1488-1497.

[6] Zelen, M. (1974) The randomization and stratification of patients to clinical trials, J. Chron. Dis. 27, 365-375