Curriculum-embedded experimentation

Steve Ritter
Carnegie Learning
sritter@carnegielearning.com

April Murphy
Carnegie Learning
amurphy@carnegielearning.com

Stephen Fancsali
Carnegie Learning
sfancsali@carnegielearning.com

ABSTRACT

We contrast two types of experiments: pull-out experiments, in which students are pulled out of their normal educational environment and curriculum-embedded experiments, in which student experience experimental conditions as a normal consequence of proceeding through a curriculum. We argue that, for practical reasons, curriculum-embedded experiments are preferred for large-scale experimentation in schools, and such experiments may also avoid issues with ecological validity. The choice to use a curriculum-embedded approach affects methods used for randomization, subject screening and the way that student experience after the experiment concludes is affected by the experiment, even for students who did not themselves participate.

Author Keywords

Field experimentation; A/B testing, group-randomization

INTRODUCTION

A large class of educational research focuses on understanding which of several different approaches to teaching content works best for students. For example, Makransky and Mayer [1] compared two versions of instruction on climate change, one using video and the other immersive augmented reality. Such experiments are typically what we might call *pull-out experiments*. The experimenter recruits student participants, assigns these participants to conditions and administers the instruction, along with pre- and post-tests and surveys designed to determine the impact of each instructional approach. Such experiments may be conducted in classrooms ("in-vivo experimentation") or in labs, but in either case, they pull students out of their normal instructional context in order to perform the experiment. Careful experimenters will attempt to time the experiment and target participants such that the experiment includes students who have not yet learned the target material but who have also learned the appropriate prerequisites so that they are ready to learn the material that is the subject of the experiment.

Now consider what happens when students are using some form of adaptive curriculum in which different students in the same class may be working on different curricular topics at any given time. In such cases, a pull-out experiment is bound to include students who are outside of the range of knowledge assumed by the experiment (that is, they have either already learned the target material or they have not learned some of the prerequisites for that material). In fact, it is common for experimentation that takes place within adaptive systems to pick a "reasonable time" to run a pull-out experiment and accept the consequences that some students may test at floor or ceiling and not contribute to the experimental findings [2].

One way to address the concern with adaptive curriculum would be to run what we refer to as *curriculum-embedded experiments*. Such experiments take place within the curriculum itself. In an adaptive curriculum, a curriculum-embedded experiment allows each student to start and complete the experiment as a normal part of the instructional process. When the reach the appropriate place in the curriculum, the experiment starts for them. The experiment runs asynchronously as each student reaches the topic that is the focus of the experiment. Pre- and post-tests are inserted within the curriculum, aligned with the experiment.

Within adaptive curricula, curriculum-embedded experiments provide a more ecologically valid approach, since each student is at the appropriate place in the instructional sequence when they provide data for the experiment. Curriculum-embedded experiments may also have logistical advantages. Given access to a large potential pool of participants, experimenters may be able to run experiments on any topic at any time during the school year. This frees experimenters to

design and deploy experiments when they are ready, without being driven by the pace of the curriculum.

While curriculum-embedded experiments are advantageous when experimenting on adaptive curricula, they are essential when conducting experiments at scale [3]. The concern with appropriately timing experiments in adaptive curricula is not really driven by adaptation but by the fact that students are experiencing educational topics asynchronously with respect to each other. Once we consider running experiments that span classrooms or even schools, districts and states, we find that, even if all students are using non-adaptive curricula, their progress through the curriculum will be asynchronous with respect to any particular topic. Some of this asynchrony has to do with the pace that teachers set for students. But curriculum sequences also differ between students, schools and districts. For example, a seventh grader in an accelerated math track may address solving linear equations with variables on both sides of the equals sign early in the school year, while a seventh grader in the standard track might encounter the same topic late in the school year. State standards may also dictate that particular topics be sequenced earlier or later in the school year, and states often differ in the grade level in which a particular topic appears. There is no hope of finding an optimal time to run the experiment for all students. The choices are to run the experiment at different times in different locations (which is logistically very complex) or to pick a common time and accept that there will be issues with students who are over- or under-prepared for the experimental material.

In contrast, the curriculum-embedded experiment scales very well. You could imagine running such an experiment with thousands or even millions of students in many classes across districts. In fact, the asynchrony of students across districts and states can be a strong advantage for an experimenter. Instead of the experimenter being driven by the school schedule to get the experiment ready and fielded before the target students are addressing the target topic, experimenters working with a widely deployed system can essentially run an experiment at any time during the school year, under the assumption that asynchronous use of curriculum across a wide range of schools and students will provide an adequate pool of subjects at any given time. Experimenters can expand enrollment in an experiment simply by expanding the amount of time that the experiment is available - if enough students haven't participated in the experiment in one month, simply leave the experiment open to enrollment for another month.

Our design of UpGrade [4] has focused on curriculum-embedded experiments, but, in some ways, we are only beginning to realize how such experiments differ from inserted experiments. In many ways, curriculum-embedded experiments resemble many clinical medical experiments, in which patients at a clinic or hospital are screened based on a protocol to determine whether they are eligible to be in the experiment and, if so, they are randomized and treated according to that random assignment.

Randomization

One important way that curriculum-embedded experiments differ from inserted experiments is in the way randomization is performed. In inserted experiments, the experimenter typically identifies the pool of students (or classes or schools) who will participate and can assign students to a condition before the experiment begins. Simple randomization, in which the experimenter randomly assigns each student to a condition is common, but more sophisticated and statistically powerful randomization techniques are also possible [5]. For example, experimenters may use paired or stratified randomization [6], in which students are grouped according to various characteristics that are assumed to be relevant to the educational outcomes. For example, students with similar socio-economic status (SES), prior academic achievement and special education status might be grouped together and then students in this group evenly distributed between conditions. This procedure ensures that each condition is similar with respect to the characteristics used to group students. Controlling such variables increases statistical power.

In a large-scale curriculum-embedded experiment, it is often not possible or practical to identify the students who would encounter the topic of interest during the experiment time window. Since participants are not known in advance, we cannot pair or stratify students in advance of the experiment to ensure that conditions are similar with respect to various characteristics. Medical clinical experiments often use a form of adaptive randomization [7] such as a biased coin design [8]. In this kind of design, the probability that a subject will be assigned to a particular condition varies so as to assign subjects to conditions such that the conditions are kept relatively similar along various characteristics.

UpGrade currently uses simple randomization to sequentially assign subjects to condition, but a planned improvement will include stratified random assignment, which will allow users to more carefully match students between conditions. We are not yet sure how common it would be for UpGrade users to be able to identify the pool of participants in an experiment and are still exploring the possibility of supporting paired or stratified randomization.

SCREENING PARTICIPANTS

As discussed earlier, a primary consideration in educational experiments is that the experiments be performed on students who have not yet been taught the material that is the subject of the experiment. It makes little sense to test the efficacy of instruction on students who have already fully learned the topic. Thus far in the discussion, we have treated the curriculum as a sequence of topics. Curriculum-embedded experiments take place at some point in that sequence. In truth, students' path through a curricular sequence is not always strictly linear. Some students may return to a topic to review it before a test, for example. A teacher may direct a student to repeat a topic if the student shows evidence of having forgotten or incompletely learned the material.

Similar to a clinical medical trial where patients may be excluded from an experiment due to prior treatment for a condition, within UpGrade, we have implemented the ability to exclude students from the experiment based on their instructional experience prior to the experiment. UpGrade has the ability to record students' coverage of curricular topics, and experimenters may specify whether experience with a particular topic or set of topics should exclude a student from the experiment. The ability to specify such rules in advance is essential for running large-scale experiments, where such decisions can not be made on a case-by-case basis.

POST-EXPERIMENT BEHAVIOR

Large-scale experiments also need to define rules for how the educational experience should progress after the experiment is completed. Consider the case where a student participated in an experiment on fraction division and received a new, experimental version of instruction on that topic. The experiment then ends, but the student goes back to review fraction division before a test. The student's expectation would probably be to receive the same approach to fraction division while reviewing the test as in primary instruction. In this way, the ability to deliver the experimental treatment to students continues, even after the experiment is no longer enrolling students. This process also has parallels in the medical literature, where treatments may continue beyond the formal end of study accrual and analysis.

One way in which educational experiments may be different from most medical studies is in group random assignment contexts [9]. Educational interventions often take place in group settings (like classrooms), and it may be desirable to provide the same intervention to all students in a class. Assignment by class helps address perceived (or real) unfairness concerns among students and may make implementation easier on teachers, who do not need to support different instructional approaches for different students in their class. In order to maintain this within-class consistency, it may be necessary to provide the experimental treatment to students who are not participating in the experiment. Consider the case where an experiment on a particular topic started and was completed before one or more students encountered the topic. If these students encounter the topic after the experiment completes, they should get assigned to the experimental condition experienced by the rest of the class, even though the experiment has been completed. Since the experiment has ended, these students would not be considered in data analysis (and may not be administered tests and surveys associated with the experiment), but their educational experience would continue as if they were in the experiment.

CONCLUSION

In many educational settings, practical considerations dictate that large-scale experiments be curriculum-embedded experiments, rather than pull-out experiments. Educational researchers maybe unfamiliar with curriculum-embedded designs and it can be difficult to understand the implications of this kind of design for randomization, screening and maintaining the ability to continue delivering experimental conditions even after the experiment has ended. The UpGrade system is designed to provide support for curriculum-embedded designs.

ACKNOWLEDGEMENTS

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305N210045 to Carnegie Learning, Inc. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

REFERENCES

[1] Makransky, G., Mayer, R.E. Benefits of Taking a Virtual Field Trip in Immersive Virtual Reality: Evidence for the Immersion Principle in Multimedia Learning. Educ Psychol Rev(2022). https://doi.org/10.1007/s10648-022-09675-4

[2] Butcher, K. R., & Aleven, V. A. (2008). Diagram Interaction during Intelligent Tutoring in Geometry: Support for Knowledge Retention and Deep Understanding. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1736- 1741). Austin, TX: Cognitive Science Society.

[3] Stamper, J.C., Lomas, D., Ching, D., Ritter, S., Koedinger, K.R, and Steinhart, J. (2012). The Rise of the Super Experiment. In *Proceedings of the 5th International Conference on Educational Data Mining (EDM 2012)*. 196-200.

[4] 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 S. E. Fancsali, S. Ritter, & A. Murphy (Eds.), *Proceedings of the First Workshop on Educational A/B Testing at Scale*. EdTech Books. https://edtechbooks.org/ab_testing_2020/upgrade_ab

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

[6] Fairhurst, C., Hewitt, C. E., & Torgerson, D. J. (2020). Using pairwise randomisation to reduce the risk of bias. *Research Methods in Medicine & Health Sciences*, 1(1), 2–6. https://doi.org/10.1177/2632084319884178

[7] Pocock, S.J. and Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics* 31(1), 103-115.

[8] Frane, J.W. A Method of Biased Coin Randomization, Its Implementation, and Its Validation. *Therapeutic Innovation and Regulatory Science* 32, 423–432 (1998). https://doi.org/10.1177/009286159803200213.

[9] Ritter, S., Murphy, A., Fancsali, S. (2020). Managing group random assignment in UpGrade. First Workshop on Educational A/B Testing at Scale. In S. E. Fancsali, S. Ritter, & A. Murphy (Eds.), *Proceedings of the First Workshop on Educational A/B Testing at Scale*. EdTech Books. https://edtechbooks.org/ab_testing_2020/managing_group

[10]