Large-scale digital experimentation in classrooms: design considerations for adaptive software

April Murphy
Carnegie Learning
amurphy@carnegielearning.com

Steve Ritter
Carnegie Learning
sritter@carnegielearning.com

Large-scale, classroom-based experiments using adaptive instructional software pose somewhat unique challenges for experimental design and deployment. One reason for this is that adaptive software allows students to advance through the curriculum at different rates and encounter content at different times, meaning that content targeted for experimentation is often reached asynchronously by students within the same classroom. In addition, many pedagogical approaches subject to experimentation will require multiple "touch points" with students throughout a course. Nimble experimental methods necessitate that experiments are able to take place at any time during the school year, so a robust experimentation system may need to be aware of the student's educational experience(s) prior to the experiment, both to permit students to be excluded from the experiment (if desired by the researcher) if they have previously been taught the target content, and to ensure that the pedagogical approach taken during the experiment is consistent with that which has been experienced by the student before and after the experimental period.

The UpGrade A/B testing platform (Ritter et al., 2020) is designed to help experimenters navigate these issues. This paper discusses several key design considerations for conducting digital experiments within adaptive educational software: ordering and sequencing, coordination of experimental activities, and exclusion criteria. Following this is an illustration of how these principles were applied in two recently-conducted large-scale experiments.

Ordering and sequencing

Educational materials often adapt a narrative form. For example, a textbook is written to be consumed from the first chapter to the last, similar to chapters in a novel or scenes in a movie. In the educational context, topics are sequenced such that they obey prerequisite relationships. Unlike narratives in other domains, though, educational sequences are often customized to match state or national standards or to satisfy instructor preferences. It is common, for example, for an instructor to decide to skip a chapter, to substitute a chapter in the textbook for one in another book or to present the chapters out of sequence. Adaptive software works the same way but at the individual level. Instructors typically have control over the sequence and inclusion of topics in the curriculum at the class and individual levels. Adaptive software may also include, omit, extend or contract the presentation of particular topics for individual students, based on its evaluation of the student's needs. As with other narratives, components of the educational narrative are designed to be encountered by students only once. Students may re-read a chapter but most do not.

Within educational software, experiments might be concerned with general features which apply across all content, such as the general UI or the presence of resources like a glossary or with activities presented in the software. Our concern here is with experiments that are linked to educational activities. In an experiment designed to assess the impact of a particular activity (like completing a book chapter), we might want to differentiate students encountering the chapter for the first time from those re-reading the chapter. We might also want to guarantee students a consistent experience with that chapter. For example, if a student re-reads a chapter (to prepare for a test, for example), the student might reasonably expect that chapter to be identical to the one they initially read, even if an experiment takes place at one encounter but not the other.

Within adaptive software, students might be completing the same activities at different times, meaning that an activity-linked experiment will be running for different students at different times, even for students in the same class. Typical "in-vivo" classroom experiments deal with this issue by disengaging students from their normal instructional context to deliver experimental interventions at a particular time window. While such approaches are advantageous in the amount of control researchers have over the study, the benefits turn into drawbacks when attempting to scale (Stamper, 2012), since carefully timing and controlling the manipulation can become untenable when conducted over hundreds or thousands of classrooms. The fixed-time-point, "pull out of context" approach is also poorly suited to adaptive learning curricula (Ritter et al., 2022). In such cases, students may have either already learned the target experimental material or have not yet mastered prerequisites for the material, leaving few participants in the ideal experimental "window". An alternative solution is to conduct a "curriculum-embedded" experiment, where students encounter the experiment within their normal instructional sequence. Students reach the target experimental topic and the experiment begins automatically, without disruption. The need to embed experiments within the curriculum becomes crucial when conducting experiments at scale across many states and districts, since students' progress is not dictated by a universal timeline.

Given these constraints, the reality that there is no optimal time to run an experiment for *all students* means that, particularly in widely-deployed adaptive instructional software, launching an experiment any time in the school year can provide useful data from *some students*. When the participant pool is thousands or millions of students, a smaller sample may still provide far more data, more quickly, and with greater statistical power than small-scale approaches.

Coordinating experimental activities

Some experiments may be designed to extend across multiple, rather than single, activities in a curriculum sequence. For example, researchers may add experimental interventions to multiple instructional modules within a unit topic. In such approaches, the elements of a student's activity sequence may be determined by condition assignment; e.g. in a curriculum sequence made up of five activities, students randomly assigned to the experimental condition would receive both Activity 2a and Activity 4a rather than Activity 2 and Activity 4 (Figure 1). We refer to the site at which content may diverge in an educational experiment (for instance, immediately after completing Activity 1) as the *experimental decision point.*

There are three options to managing student condition assignment when experiments involve multiple interventions. One approach is to have randomization occur at the start of the experiment (e.g. prior to Activity 1) with condition assignment consistent for each student across all coordinated activity elements. For example, if Activity 2 has to do with adding fractions and activity 4 has to do with subtracting fractions and there is some commonality in the approach to fractions taken in the experimental versions, we might want to ensure that assignment to experimental activities is coordinated across activities. In UpGrade, we call these *coordinated decision points*. A second option is for the experimenter to allow randomization to occur independently at each point (e.g., once at the end of Activity 1, and once at the end of Activity 3), with participants potentially receiving different condition assignments at each decision point, creating a "dosage effect", with each student eligible to receive an experimental condition 0, 1, or 2 times in the illustrated example sequences. This might be the case, for example, if the experiment were testing the use of worked examples in fraction instruction. A third possibility is to control the dosage in a within-subjects experiment. We might want to ensure that students get a worked example either in Activity 2 or in Activity 4 but not in both. Note that the decision about which of these three ways to design the experiment has to do with the goals of the experiment and the particular educational content. Our goal is UpGrade is to allow experimenters to specify how to manage these issues, not to try and advise the experimenter on which to choose.

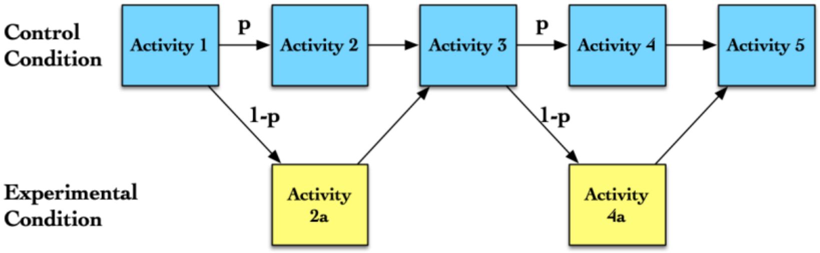


Fig. 1: An activity sequence with two coordinated decision points

Exclusion reasoning

Another challenge presented by adaptive educational software is how to take into account students' prior learning experiences. As noted earlier, embedding experiments in the curriculum is an approach that ensures students reach the experimental content at a pre-defined point in the sequence, but students sometimes repeat an activity. Many experiments will want to distinguish between first and subsequent encounters of an activity. In many cases, prior experience with an activity may disqualify a student from the experiment, so it may be important to track students' experiences with activities, even before the experiment begins. For example, in the coordinated activities case, a student who completes Activity 2 prior to the experiment starting might be excluded from the experiment so that Activity 4 will match the student's experience with Activity 2.

In UpGrade, we have implemented the ability to automatically exclude students from an experiment based on their prior learning experiences, and allow experimenters to determine whether, as in the "dosage effect", a student assigned to the experimental condition must experience *all* possible activities associated with that condition to be included in the experiment.

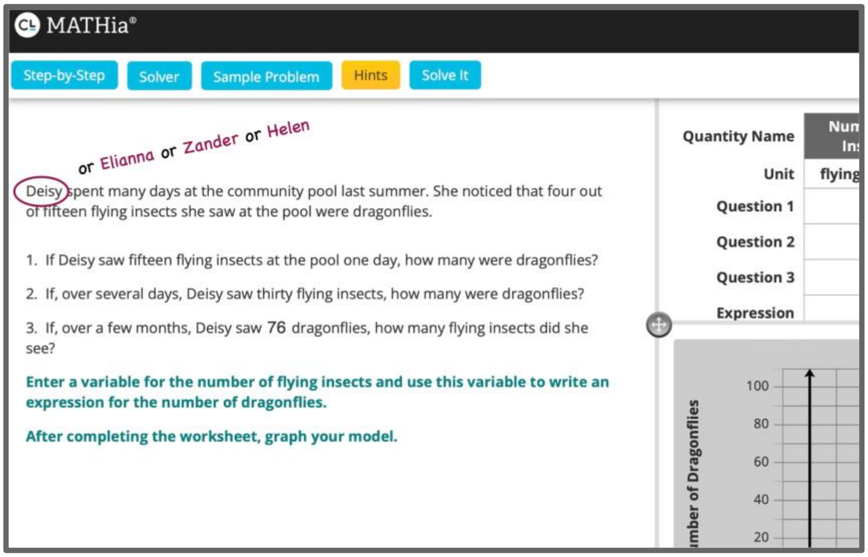
Example experiment designs

These considerations are best illustrated with respect to specific experiments. Here, we describe two experiments that we ran in MATHia (Ritter et al., 2007), our adaptive software for teaching mathematics in middle and high schools. Within MATHia, students are assigned a sequence of math topics, called workspaces, according to their grade level and course. They can progress through workspaces at their own pace. Most workspaces present a variable number of problems to each student, depending on the student's ability to demonstrate mastery of problems relevant to the topic addressed in the workspace.

Personalization Experiment

Based on increasing recognition that students' sense of belonging in school can impact their academic achievement (Walton & Cohen, 2007), we were interested in whether this sense would increase among underrepresented student populations if they recognized that their math problems were personalized for students like them. To test this hypothesis, we conducted a large-scale experiment across three school districts. In the experiment, we generated word problems where, in the experimental condition, the (first) names of people referenced were drawn from a "localized" list of names taken from the student's community, rather than from a standard, nationally-normed list. For example, in the control condition, students completed a word problem in which "Helen" is spending money at a certain rate, while in other school districts, the character name was "Deisy," "Elianna" or "Zander."

Not every workspace in MATHia involves word problems and, to maximize impact, we wanted students in the experimental condition to experience multiple workspaces with localized names. With respect to Figure 1, this experiment design is one in which Activities 1, 3 and 5 are *not* word problem workspaces and so are identical for all students. Activities 2 and 4 involve word problems, and students in the experimental condition would receive word problems with localized names instead of the standard set (in the actual experiment, we localized 10 workspaces, with 60-100 workspaces typical of a full-year curriculum). The experiment started in March, so some students had already completed some of the target workspaces prior to the start of the experiment (meaning that those students saw the default names). The study was conducted across multiple schools in districts, so there was variation in workspace ordering (curriculum sequences can differ across states and individual students, and state standards often dictate when topics are sequenced in the school year).



Deisy spent many days at the community pool last summer. She noticed that four out of fifteen flying insects she saw at the pool were dragonflies.

1. If Deisy saw fifteen flying insects at the pool one day, how many were dragonflies?
2. If, over several days, Deisy saw thirty flying insects, how many were dragonflies?
3. If, over a few months, Deisy saw 76 dragonflies, how many flying insects did she see?

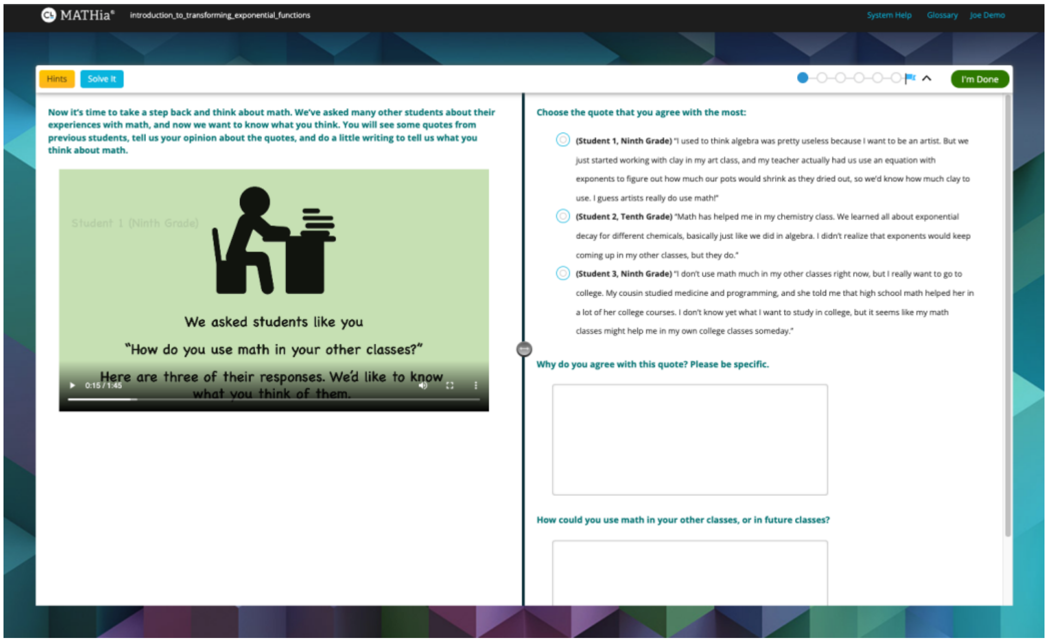
Enter a variable for the number of flying insects and use this variable to write an expression for the number of dragonflies.

After completing the worksheet, graph your model.

Fig. 2: Illustration of one problem from the "personalization" workspace. Names used in the problem matched common names in the student's community.

We wanted to maximize the difference between the conditions, so we used coordinated decision points to ensure that students maintained a consistent condition (i.e., control or experimental) over time and workspaces. To maximize participation, we did not exclude students who had previously encountered one of the workspaces, since we did not believe that returning to a workspace and have it personalized (if randomized to a condition) would not pose any conflict with a previous encounter with a non-personalized version of the workspace. In a different context (such as if the pedagogical approach taken in one workspace depends on the approach taken in a prior workspace), we might exclude students who had previously completed one of the target workspaces; UpGrade allows the experimenter to make this decision.

Utility-value Experiment



Now it's time to take a step back and think about math. We've asked many other students about their experiences with math, and now we want to know what you think. You will see some quotes from previous students, tell us your opinion about the quotes, and do a little writing to tell us what you think about math.

We asked students like you

"How do you use math in your other classes?"

Here are three of their responses. We'd like to know what you think of them.

Choose the quote that you agree with the most:

○ **(Student 1, Ninth Grade)** "I used to think algebra was pretty useless because I want to be an artist. But we just started working with clay in my art class, and my teacher actually had us use an equation with exponents to figure out how much our pots would shrink as they dried out, so we'd know how much clay to use. I guess artists really do use math!"

○ **(Student 2, Tenth Grade)** "Math has helped me in my chemistry class. We learned all about exponential decay for different chemicals, basically just like we did in algebra. I didn't realize that exponents would keep coming up in my other classes, but they do."

○ **(Student 3, Ninth Grade)** "I don't use math much in my other classes right now, but I really want to go to college. My cousin studied medicine and programming, and she told me that high school math helped her in a lot of her college courses. I don't know yet what I want to study in college, but it seems like my math classes might help me in my own college classes someday."

Why do you agree with this quote? Please be specific.

How could you use math in your other classes, or in future classes?

Fig. 3: Illustration of the manipulation in the utility-value experiment.

In a second large-scale experiment, we focused on utility-value interventions (Harackiewicz et al, 2016), and whether increasing students' perception of the value of mathematics in their everyday lives could affect their achievement and attitudes towards mathematics . In this experiment, conducted with over 13,000 students in over 500 schools, students in the experimental condition saw videos which presented students talking about real-world examples of the mathematics topic they were learning (for example "going viral" on Instagram in a workspace about exponential growth). Within this experiment, we randomize at each decision point, rather than using coordinated decision points. As a result, each student received a particular "dosage" of utility-value activities. The experiment encompassed 8 workspaces, six that presented utility-value videos and 2 used for the pre- and post-tests (which were identical for all students). For individual students, encountering a utility-value workspace depended on the curriculum that teachers assigned to the student as well as the extent to which the student's work in MATHia overlapped with the experiment window. As a result, individual students might encounter between 0 and 6 utility-value videos. As in the belonging experiment, we configured UpGrade such that prior experience with any of the workspaces presenting a utility-value video would not disqualify the student from the experiment. However, workspaces used for the pre- and post-test were marked as disqualifying; students who had done one of those workspaces prior to the experiment were excluded from the study.

Conclusion

The combination of narrative structure in course materials, flexibility in the presentation of those materials and the use of adaptive software requires careful consideration of how the experiment window interacts with the way students encounter content. In particular, experimenters need to think about whether instructional activities that take place at different times need to be coordinated and about whether prior experience with particular content should be disqualifying for potential participants. These requirements appear common in educational contexts but likely uncommon in other contexts. We have developed the UpGrade system to help experimenters manage these factors, and we have illustrated the utility of these features with respect to two recently fielded experiments.

Acknowledgement

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

Harackiewicz, J. M., Canning, E. A., Tibbetts, Y., Priniski, S. J., & Hyde, J. S. (2016). Closing achievement gaps with a utility-value intervention: Disentangling race and social class. *Journal of personality and social psychology*, *111*(5), 745.

Ritter, S., Murphy, A., Fancsali, S.E. (2022). Curriculum-embedded experimentation. *Proceedings of the Third Workshop on A/B Testing and Platform-Enabled Research (at Learning @ Scale 2022).*

Ritter, S., Murphy, A., Fancsali, S.E., Fitkariwala, V., Patel, N., & Lomas, J.D. (2020). UpGrade: An open source tool to support A/B testing in educational software. *Proceedings of the First Workshop on Educational A/B Testing at Scale (at Learning @ Scale 2020).*

Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive Tutor: Applied research in mathematics education. *Psychonomic bulletin & review*, *14*(2), 249-255.

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.*

Walton, G. M., & Cohen, G. L. (2007). A question of belonging: race, social fit, and achievement. *Journal of personality and social psychology*, *92*(1), 82.