Evaluating Educational Interventions at Scale

Rakesh Agrawal
Microsoft Research

M. Hanif Jhaveri
Stanford University

Krishnaram Kenthapadi
Microsoft Research

INTRODUCTION

Education and learning are currently undergoing transformative changes due to the emergence of tablet devices, cloud computing, and abundant online content. These trends present opportunities to transform traditional paper-based textbooks into tablet-based electronic textbooks, and to further enrich the educational experience by augmenting them with relevant supplementary materials [1]. A natural question is whether this educational intervention, namely, enriching textbooks with relevant web articles, images and videos, is effective. It turns out that designing an experiment at scale for this purpose is nontrivial. We report on progress in designing and carrying out such an experiment.

CLASSICAL APPROACH

Randomized control trial is often deemed the gold standard for impact evaluation [2]. Its key feature is that the study subjects are randomly allocated to receive one or other of the alternative treatments under study. Those in the treatment group are compared to those who were randomly assigned to the control group – those who did not receive the intervention. Because members of the groups (treatment and control) do not differ systematically at the outset of the experiment, any difference that subsequently arises between them can be attributed to the treatment rather than to other factors. The post-intervention results analysis can lead to the refinement of the intervention and the randomized trial is repeated with the revised intervention (Figure 1).

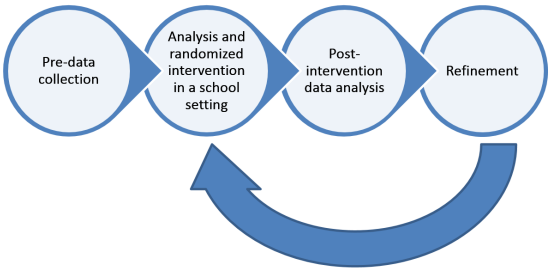

Figure 1. Classical Evaluation Methodology

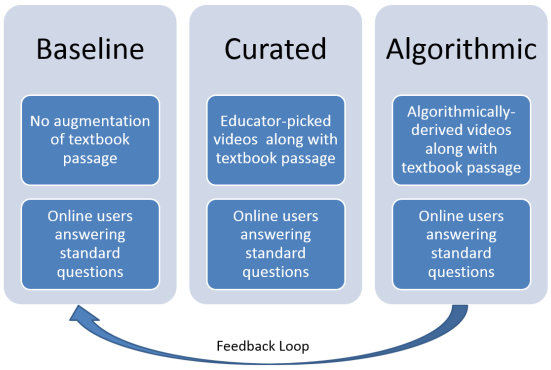

Figure 2. Evaluation Design for Textbook Augmentation

However, the following issues arise immediately in applying randomized control trials to the task of determining whether a supplementary material helps improve the understanding of a textbook passage, particularly when the educational material will be used across geographies:

- Intervention in classrooms has limited sequencing since several months are needed to get any feedback in typical school settings, and moreover, is very expensive.
- It is very difficult to ensure the requisite diversity, or even to get more than one classroom at a time.
- Interventions that assume natural progressions in the building-block technologies (*e.g.*, reliable broadband internet access) and are designed for deployment in future (say, 2 to 3 years from now) are difficult to study.
- Learning involves interactions between students, teachers, and other stakeholders and thus, is not an isolated experience that can be measured, and hence separating the effect of an augmentation from a plethora of other variables is hard to achieve.

SCALABLE DESIGN

We now present our online evaluation platform under development that leverages users world-wide to carry out experiments at scale to study the effectiveness of enriching electronic textbooks with educational videos (Figure 2). The basic ingredients of our design are:

1. Baseline: Online users would be presented with the textbook passage without any augmentation.
2. Curated: Online users would be presented educator curated videos along with the textbook passage.
3. Algorithmic: Online users would be presented videos obtained algorithmically (adapting techniques proposed in [1]) along with the textbook passage.

In all three cases, the users are required to answer questions that test knowledge of the textbook passage. In the baseline experiment, we also ask the users whether they would find it useful to have educational videos in addition to the textbook

L@S'14, March 4–5, 2014, Atlanta, Georgia, USA.
ACM 978-1-4503-2669-8/14/03.
http://dx.doi.org/10.1145/2556325.2567884

passage. Through these experiments, our goal is to not only understand whether educator curated videos can help improve the performance of the user, but also to iteratively refine the algorithmic techniques to get closer to the performance obtained with the curated videos. Our design is inspired by approaches focused on understanding networks as opposed to isolated variables (*e.g.*, [3]).

IMPLEMENTATION

We present different implementation decisions, taking into account three broad dimensions: academic considerations (factoring in the rich education literature as well as recent work on online platforms), design considerations (our design goals), and iterations (based on our trials and anecdotal evidence).

1. Platform selection: We chose to use Amazon Mechanical Turk platform since this platform has been sufficiently vetted by the academic community (*e.g.*, [4, 5]). In particular, this platform has been shown to be fairly reliable, flexible, and geographically diverse, and suitable as a proxy to real world interactions. Alternate approaches such as solicitation of users through online lists/ads are hard to scale, and hence removing selection bias becomes harder.

2. Textbook passage selection: We selected a corpus of textbooks spanning different subjects (physics, chemistry, biology, economics), difficulty level (9^{th} grade to college level), and geographies (CK-12 books (USA) and NCERT textbooks (India)). We chose a set of nine passages from seven different textbooks, and asked teachers to generate ten questions, and a set of curated videos for each passage. Since many studies have shown that task lengths of 60 minutes or less are desirable in online platforms, we carefully arrived at the appropriate passage lengths, number of videos shown, and number of questions to be answered.

3. Educator selection: We chose teachers representing five large US states, balancing two key goals. We desired maximum variation of experiences across students in terms of their ethnic and socio-economic background and resource utilization, while at the same time, we ensured that the teachers had comfort and experience with using educational videos in existing lessons.

4. Curation process: Five educators were asked to select the questions (that could be answered by reading just the textbook), and a different set of five educators were asked to curate relevant videos for augmentation, so that there is no bias between the two processes.

5. Design of HIT (human intelligence task): We designed the HIT so that the entire functionality is built into the task, and used very basic web tools so that judges across different economic backgrounds are likely to have very similar experience with our task. We further benchmarked the performance with students at a US university to ensure that the task was not too difficult. We ensured that no one could participate more than once, and included honeypots to prune bad participants. We also carefully monitored to weed out participants who did not follow instructions, or spent very little time on the task. Based on several trials, we arrived at the rate of USD $2.50 per hour that attracted the most desirable participants. They often provided the optional feedback, for example, expressing their hope that their participation would indeed help future students. We could not attract quality participants below this rate. With higher rates, we were attracting participants who just wanted to earn quick money; in fact, they did not provide any feedback and rushed to complete the HIT, missing honeypot questions in the process.

6. Demographics: We conducted trials across two geographies (USA and India), with 100 users per trial. We collected demographic data to ensure that the distribution of the judges matches the overall target distribution.

7. Selecting participants: Given the relatively large cognitive complexity of our task (requires understanding of the context of the textbook material as well as the video), we wanted to only include judges who had the prerequisite analytical and reading comprehension abilities. We included a set of five questions pertaining to analytical and reading comprehension abilities, and excluded judges who answered fewer than two of the five questions correctly.

PROGRESS REPORT

Our initial results suggest that the videos would indeed be helpful for enhancing the experience of learning from the textbooks. Of all the participants, 65% of them said that it will be helpful to have videos in addition to the textbook passage. We observed that the corresponding percent was higher for Indian participants (73%), compared to US participants (57%). A plausible explanation is that English is not the native language for most Indian participants while the textbooks are in English, and hence these participants are likely to benefit more from having explanatory videos on the subject material. We were initially skeptical whether a task with relatively large cognitive complexity such as ours could even be performed over the Mechanical Turk platform. We were pleasantly surprised to not only find many takers, but also to observe that 60% of the prequalification questions were answered correctly on average. We are currently in the process of performing extensive trials, towards measuring the performance of the algorithmic approach, and iteratively refining the underlying techniques.

REFERENCES

1. Agrawal, R., Christoforaki, M., Gollapudi, S., Kannan, A., Kenthapadi, K., and Swaminathan, A. Mining videos from the web for electronic textbooks. Tech. Rep. MSR-TR-2014-5, Microsoft Research, 2014.
2. Glennerster, R., and Takavarasha, K. *Running Randomized Evaluations: A Practical Guide*. Princeton University Press, 2013.
3. Hidalgo, C. A., and Hausmann, R. A network view of economic development. *Developing alternatives 12*, 1 (2008).
4. Mason, W., and Suri, S. Conducting behavioral research on Amazon's mechanical turk. *Behavior research methods 44*, 1 (2012).
5. Paolacci, G., Chandler, J., and Ipeirotis, P. Running experiments on Amazon mechanical turk. *Judgment and Decision Making 5*, 5 (2010).