A Framework for Evaluating the Impact of Production Quality on Coding Demonstration Videos

Malhar Pandya, Aditya Kulkarni, Brian Harrington
University of Toronto
malhar.pandya@mail.utoronto.ca, aditya.kulkarni@mail.utoronto.ca, brian.harrington@utoronto.ca

UNIVERSITY OF TORONTO SCARBOROUGH

Abstract

This work reports on a pilot study that examines the impact of the production quality of coding demonstration videos on student learning. With an increase in demand for asynchronous content, many instructors are finding the scripting and development of coding demonstration videos to be cumbersome and time-consuming. However, it is an open question whether it is necessary to carefully script and produce videos for achieving better outcomes.

In this study, participants watched coding demonstration videos, and the impact was measured on their responses to attitude surveys and content questions. Students were randomly assigned to watch either high-production quality videos which were carefully scripted and edited, or low-production quality videos which were created on the fly with no preparation time or post-video editing.

In this pilot study, no statistical significance in test performance was found based on the production quality of the videos. There was an impact on responses to a single question on the CAS[1].

This work develops a methodology for further evaluation of the impact of production quality, and aims to provide guidance to practitioners regarding the amount of time and effort required to optimize learning outcomes.

Results

- Average programming question grades improved in both participant groups, with no significant difference between them.
- One CAS survey question showed a notable change between groups: "**After I study a topic in computer science and feel that I understand it, I have difficulty solving problems on the same topic.**"
- Shifts in responses (post-intervention minus pre-intervention) were calculated:
  - Control: Negative shift (**−0.4375**)
  - Experimental: Positive shift (**+0.4138**).
- Applying a two-sided Mann-Whitney *U* Test on these shifts revealed a Bonferroni corrected p-value below **0.022**, with an effect size of approximately **0.282**.

Methodology

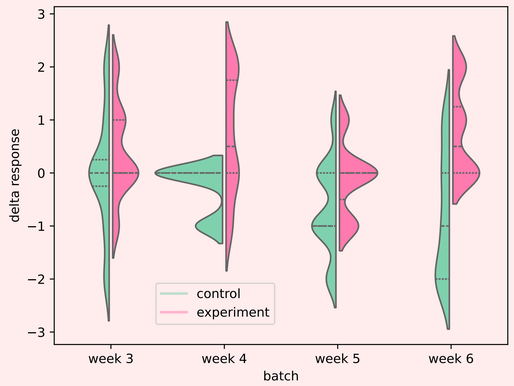
- Conducted in an introductory programming course for non-majors that used Python, across the four central weeks [3-6] of the semester.
- Each week's topic had been introduced in the preceding week's lecture, but students had not been evaluated on them yet.
- 61 Students participated in an asynchronous session, in exchange for a bonus mark in the course. They logged on to an external online platform, choosing their own two-hour window within an assigned two-day period.
- The session consisted of five timed sections:
  1. A pre-intervention survey (15 min)
  2. A pre-intervention coding question (25 min)
  3. A coding demonstration video (30 min)
  4. A post-intervention coding question (25 min)
  5. A post-intervention survey (15 min)
- Both surveys were identical subsets of questions from the Computing Attitude Survey (CAS)[1].
- The coding questions and demonstration video were on the same topic(s) and of similar difficulty.
- Participants were manually graded on their coding submissions based on objective learning outcomes described in the course outline.
- The coding demonstration video began with a problem statement, followed by the identification of solution steps and their implementation, then concluded with running the resulting program.
- Full control over video navigation was provided.
- There were two treatments for the coding demonstration video based on production quality.
- High Production Quality [HPQ]:
  - Received by the control group.
  - These videos were produced in a professional setting, with a script, and procedures outlined in *Stephenson*[2].
  - The scripts were refined by the course instructor to ensure accuracy and clarity.
  - These videos also featured a separately recorded intro and outro section.
- Low Production Quality [LPQ]:
  - Received by the experimental group.
  - These videos were produced by experienced TAs, but they did not know the problem statement prior to the recording.
  - Each presenter was given a 'practice run' to understand the question and devise a solution. Their second attempt was then used with no post-processing.
- More details are available at https://github.com/cms-urg/Video-Styles

Analysis of Code Submissions

- Missing code submissions were given a grade of 0.
- If both pre/post coding submissions were missing, the response was excluded from analysis.
- The grades showed an improvement after the intervention, but no significant difference across groups could be identified.
- The observed increment was primarily driven by responses from week 5.
- Time spent watching the intervention video was also analyzed; no significant differences across the groups could be identified.
- Change in grades across the intervention (post-intervention minus pre-intervention) showed differing correlations with intervention duration:
  - Control: Negative correlation (≈ −0.18)
  - Experimental: Positive correlation (≈ +0.165)

Analysis of Survey Responses

- An attention check question was included in the survey; responses failing to answer it correctly were excluded.
- Remaining responses required no imputation.
- Due to the ordinal nature of the response, we used a two-sample Mann-Whitney *U* Test[3].
  - Null hypothesis: The two sampled populations are stochastically equal.
  - Alternative hypothesis: Samples come from populations with different distributions.
- The Likert scale responses were transformed to a numerical scale [1-5], and the change in survey responses (post-intervention minus pre-intervention) was computed.
- Only one survey question showed a noticeable difference across groups. This difference was consistent across all four weeks.
- For this question, the control group's post-intervention response was lower than their pre-intervention response, while the experimental group showed the opposite effect.
- Prior to testing if this difference was significant, the similarity of pre-intervention responses between the groups was assessed. A p-value of nearly 0.7 was observed, indicating no significant difference in underlying distributions.
- However, testing the change in responses for the two groups yielded a p-value below **0.002**, and an effect size of about **0.282**.
- A Bonferroni correction was applied, giving a final p-value below **0.022**, indicating significant differences in how the the treatment affected responses to this question.



Smooth violin plots of the survey response 'delta', split by group (lines indicate quartiles)

Future Work

- Many factors that could influence student learning and self perception remain unexplored. We list a few potential areas below:
- Investigating if presenter characteristics such as ethnicity, gender, age and verbal accents impact student engagement and comprehension.
- Considering whether production environment factors such as background noise, audio/video quality, and IDE themes impact learning outcomes and engagement.
- Assessing student engagement by expanding the survey questions or monitoring additional metrics during the intervention.
- Evaluating differences in ling-term retention across the two treatments.

References

[1] Allison Elliott Tew, Brian Dorn, and Oliver Schneider. 2012. Toward a validated computing attitudes survey. In *Proceedings of the Ninth Annual International Conference on International Computing Education Research (ICER '12)*. Association for Computing Machinery, New York, NY, USA, 135–142.

[2] Ben Stephenson. 2019. Coding Demonstration Videos for CS1. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education (SIGCSE '19)*. Association for Computing Machinery, New York, NY, USA, 105–111.

[3] Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics, 18*(1), 50–60. https://doi.org/10.1214/aoms/1177730491