Slide-viewing patterns among students

**Hypothesis:** Students tend to attempt problems without viewing the prerequisite slides on Grok Learning.

**Question:** To what extent do students attempt problems without viewing the prerequisite slides?

At the highest level, we are looking for a fraction (across all problems, students) of those attempts for which the prerequisite slides were not visited. Then, we dig deeper – what does it tell us? We look at across all problems and see if there’s a distribution across all problems and students. Alternatively, we can look at students and see how they’re different (e.g. demographic, performance, age).

What students tend to do – and how consistent is that, does the strategy change over time?

See what we can answer on this particular question (dig deeper). If we think clustering is a good approach, then we do that. Start off with the basic data first.

**Data integration:** We could look at grouping students by their institution. The educational context is important. If someone is being tested and there’s a lot of time pressure, then students in that class tend to rush through their materials – we can see that there are good and bad institutions in that respect.

**Hypothetical model:** Does the strategy change over the course of the student’s time on Grok? Can we connect the first course with the second one? Student ID, though anonymised, is consistent across all courses. We could look at other courses that they did. We could drill down across just the students who have done several different courses and look at whether the strategy changed from year to year.

**To do:** Work out if there are enough students who have done multiple courses.

**Note:** Some students do two courses at the same time.

Newbies are different to the beginner stream. There’s newbies, **beginners**, beginners-blockly, **intermediate, advanced** (only a few 100 doing it each year). It would be useful to compare beginners and intermediate.

**Larger question**: How do we relate this to outcomes? We can look at assessed metrics: time it took to past, how many submissions it took to past, whether they did the next problem afterwards. This would show that these strategies are good – we should encourage people to do the reading beforehand.

**Submit literature review soon:** Start working on this NOW. We can look at if someone else has done things in this area. We should look at: Australian Computer Education Conference, Educational Data Mining Conference, Artificial Intelligence in Education, International Journal of Artificial Intelligence in Education. The model used in Sophia’s paper is related to educational psychology that deals with how people approach education. Although learning styles may be out of favour, there is discussion around surface learning and deep learning.

Jumping the slides and going straight to the problem is an example of a surface learning approach – do the bare minimum – do work to get past things/spam submissions. You can probably argue that a lot of students don’t take a deep-learning approach, even good students when they are faced with time pressures, will be forced to go through surface learning approaches.

It would be interesting to see what kind of surface learning approaches do we see? There are distractions and time pressures. Lazy surface learning: Given an hour of time and doesn’t use it. ‘Good surface learning’: Able student who has less time to do the task, they’re just trying to do whatever they can in the time available.

Often, surface learning approaches take longer than the case where the student had done the task properly. People could end up making their lives harder by trying to avoid doing work.

It would be interesting to quantify how many people are like that within this group. Do we see differences based on age and gender? Do we see people learning from their mistakes, and disavowing that approach later on? Most importantly, do we see the impact? Does a surface learning approach lead to poorer outcomes in the end?

Some people lose motivation when they see a hard problem, and then procrastinate.

Some courses are more interactive. It would be interesting to see whether those kinds of slides engage students better. People have a desire to make progress.

The questions we are asking here are more from the educational research point of view. One of the things that is not clear is what is the right technique – in EDM, there’s an incentive to use more data mining techniques rather than more basic analysis. This does dictate where you would publish. Some people care more about techniques than findings.

Big Five Personalities

Choose 2-3 topics from the above and see what you can find from those journals. Look at other journals too. You need to find the important people/group. Find a small number of important and good quality papers. When you’re looking for papers, look at the final authors/supervisor (often you have research students).

Simon

Raymond Lister

Jaqui Whalley

There are different generations. Simon and Raymond are from the 1990s and early 2000s. These days you have Paul Denny, Andrew Luxton Reilley – these are the Australasian people. There’s a lot of research in the US – Amy Jo Ko, Mark Guzdial, BlueJ at KCL. They tend to publish ITiCSE, ICER, ACE, KOLI, SIGCSE-TS as well as journals. There is a much broader education area – Frontiers in Education. These topics of surface learning are not unique to just CS. They are much more widely known about.

In terms of navigating this, you start in the immediate vicinity of some of those more prominent conferences. Then branch out, look at where they’re referring to. Google Scholar is great. Look at citation counts.

**Data:** The dataset that was used was the event data containing all events that have occurred on the Grok Platform. The event data was initially stored as JSON (event.json), before being streamlined and converted into a CSV file to allow for greater interpretability. The resulting CSV file (event.csv) has five column names: event\_name, created\_at, user\_id, event\_data and condition. For the purposes of this investigation, only two events were used: problem\_run indicating that the student ran the problem and slide\_view indicating that the student viewed the slide.

**Process:** Using event.csv, the problem name was extracted from the event\_data column using string processing. Then, a script was written to iterate through the dataset (first 1000, 10000 and 100,000 rows). While iterating through the dataset, the program would check if a student had attempted a problem (indicated by the occurrence of a problem\_run event) before they had viewed the accompanying slides.

**Results**

|  |  |  |
| --- | --- | --- |
| Sample size (first n rows) | Number of instances where students attempted the problem before viewing the slide | Percentage of total sample size |
| 1000 | 15 | 1.5 |
| 10000 | 604 | 6.0 |
| 100000 | 11569 | 11.569 |

|  |
| --- |
| **# INITIALISATION**  CURRENT\_STUDENT = 0  CURRENT\_PROBLEM = 0  RUN\_PROBLEM = FALSE  VIEWEDS\_SLIDE = FALSE  COUNT = 0    **ITERATE THROUGH DATAFRAME:**  **# UPDATE THE CURRENT STUDENT**  IF USER\_ID != CURRENT\_STUDENT:  CURRENT\_STUDENT = USER\_ID  RUN\_PROBLEM = FALSE  VIEWED\_SLIDES = FALSE    **# UPDATE THE CURRENT PROBLEM**  IF PROBLEM\_NAME != CURRENT\_PROBLEM  CURRENT\_PROBLEM = PROBLEM\_NAME  RUN\_PROBLEM = FALSE  FIRST\_TIME = TRUE  VIEWED\_SLIDES = FALSE  **# RAN THE PROBLEM**  IF EVENT IS RUN\_PROBLEM  RUN\_PROBLEM = TRUE    **# VIEWED A SLIDE**  ELSE IF EVENT IS SLIDE\_VIEW  VIEWED\_SLIDES = TRUE    IF RUN\_PROBLEM AND NOT VIEWED\_SLIDES  COUNT += 1  RUN\_PROBLEM = FALSE  FIRST\_TIME = FALSE |

**In terms of making it scalable, you cannot use Pandas. Pandas doesn’t have a stream processing mode. Pandas must read in the entire dataset, and sometimes have a second copy, which incurs greater cost. For the event.csv file, what you really need is to use the CSV reader object, and then iterate through each line and process it that way – you load it up until you get a new line.**

**Filtering**

**Can we take advantage of any ordering**

**In principle, we filter based on URL to make sure we’re on the right course**

**You can maintain a dictionary of all the students – then for each problem/module – we can have the timestamp of the different visits – store a list of which problems have not been visited**

**Don’t get too bogged down in class structures – we already have dictionaries**

**If you did it by student – if student ID was the key, then we can have nested dictionaries – quite quick**

**Student as Key**

**{Student Key: {Problem Label: attributes of what happens there}}**

**We want to avoid getting bogged down in reinvention – in terms of reportable objectives, it’s more what you do with the data afterwards – we want to get to that point. Don’t get much recognition for doing this stuff.**

Ultimately,

**To dos:**

1. Extend the analysis
   1. Look at how long students spend on a given slide
   2. Look at how many students viewed the slides in order
   3. How many back and forth views of slides
   4. Most viewed slide
   5. Longest viewing time
   6. Time between slides
   7. Time to view all slides
2. Organise/pre-process in a more organised way to allow for easier analysis
   1. Classes for students and problems
3. Integrating datasets

**Reviewer’s feedback:**

1. One reviewer rejected the paper due to not believing in learning styles. Sophia shows attributes that seemed to align well with what was described in that model. However, some people take issue with issue with the model. Some people will just dismiss things with no further consideration.
2. Time spent on each slide
3. Number of times each slide has been viewed
4. Percentage of slides viewed by each student
5. Most viewed slide in each problem, and the entire course
6. Time taken to view all slides

Extension:

1. After submitting multiple slides, do people go back to view the slides?
2. Percentage of time spent on slides vs percentage of time spent on problem
3. Slide in-runs and problems