International Society of the Learning Sciences

Effort and Struggles in the Data: How do Low and High Achieving Students Use an Online Textbook in an Introductory STEM Course?

Gahyun Sung, Harvard Graduate School of Education, gsung@g.harvard.edu
Stephanie Yang, Harvard Graduate School of Education, szhang@g.harvard.edu
Mary C. Tucker, University of California, Los Angeles, maryctucker@ucla.edu
Bertrand Schneider, Harvard Graduate School of Education, bertrand_schneider@g.harvard.edu

Abstract: How students engage with the course material outside of in-class time is often an invisible, but powerful indicator of their motivation, belief, and subsequent achievement. In the current paper, we track how high and low-achieving students in an introductory data science course engage differently with an online textbook. Using log data, we create proxies for the level and type of effort students put into the coursework, and the level of frustration and struggling they experience throughout the semester. Statistical analysis and side-by-side visualizations reveal an increasingly divergent pattern of longer struggles and less effort from low-achieving learners, and also reveals the existence of a watershed week where this divergence becomes especially pronounced. We conclude with discussions on how effort and struggling metrics could be used to support learning and teaching.

Introduction

Large STEM introductory classes have long been of interest to the learning sciences community due to their gatekeeping role in STEM. Some students will gain confidence and foundational skills, take advanced courses, and go on to build a successful career in STEM fields. Others will come out of this first class with the conviction that they are not "math people", thereby closing the door on future STEM learning. The leaky pipeline phenomenon is often cited as a key mechanism of the persistent underrepresentation of women and minorities in STEM fields (Burke, Mattis & Elgar, 2007), and a hurdle to overcome if we are to meet the increasing demand for STEM professionals in today's society.

Beneath the most salient explanations for the causes such as demographics and grades (e.g., Belser et al., 2018), the lived experience of a struggling student holds rich information and opportunities for just-in-time interventions. How students come to classes, engage with the course material, prepare for tests, or spend time doing assignments will differ systematically by their underlying motivation and beliefs that evolve throughout the course. While an instructor cannot realistically track individual student states in large classes, new technologies for collecting and analyzing process data offer opportunities to continuously keep watch over how well students are doing in the course, much like a caring, expert instructor might in small group settings.
To gain actionable insights for learning and teaching, process data can be converted into theory-based, "glass box" metrics that model the relationship between student behavior and outcomes. Such metrics are human interpretable, allowing students and instructors to more intuitively understand how predictions are made, where problems are located, and where there may be errors in machine measurement and judgement. Conversely, black box algorithms that detect at-risk students with opaque processes are often criticized for their lack of insight on what might be effective remedial actions (e.g., Essa & Ayad, 2012), as well as on the risk of users categorically accepting its outputs, which can do more harm than good in the case of inaccurate profiles or false negatives (Kitto & Knight, 2019; Gillani et al., 2021). Similarly, relying on predetermined factors (e.g., demographics, prior grades) or interim student achievement for predictions is for the most part unhelpful in understanding which aspects of student behavior need attention, and run the risk of reinforcing stereotypes.

With these motivations in mind, the current paper uses data from an introductory STEM course to design interpretable behavior metrics for the end goal of informing instructors and learners when and how to improve the learning experience. In particular, for our original motivation of understanding the experiences of struggling students outside the classroom, we focus on metrics that can serve as proxies of the amount and type of effort made for the course, as well as the amount of struggling they are going through. Using these metrics, we ask whether 1) high-achieving and low-achieving students, divided by performance in quizzes, show different levels of stress and types of effort over time, as well as 2) whether we can identify watershed weeks where all students seem to be struggling, based on the metrics. We conclude with discussions on how student behavior metrics could be used to improve learning and teaching.

Methods

Setting

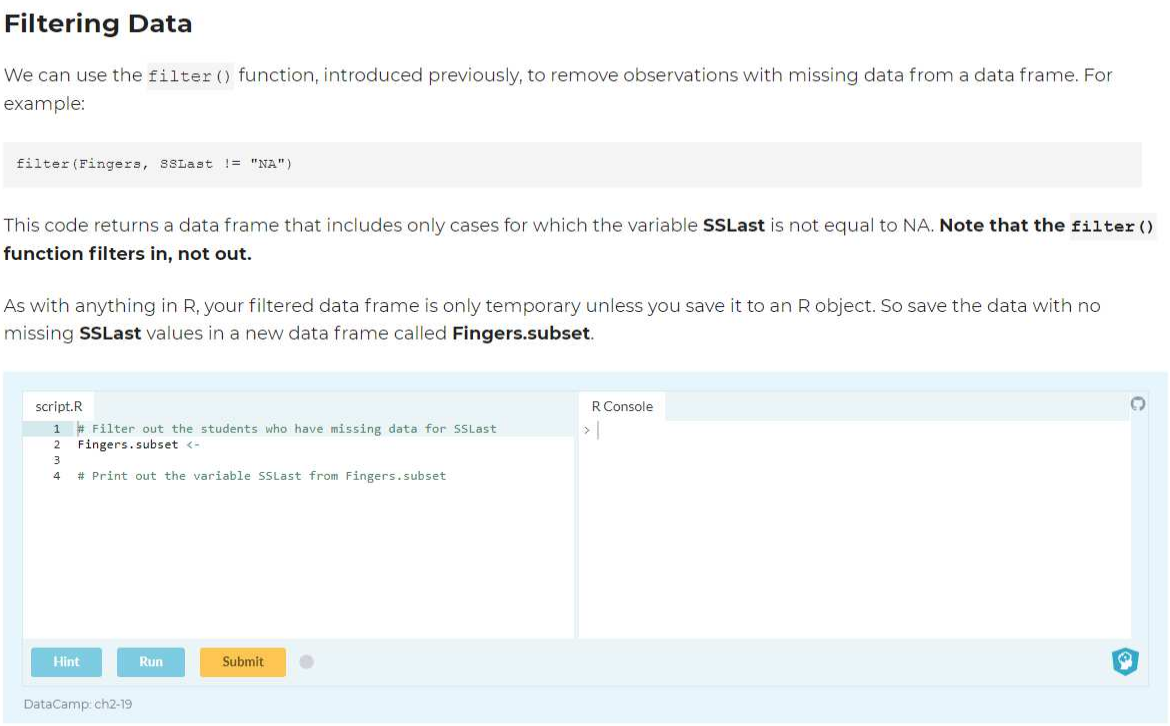
The current paper looks at one promising setting for using log data to gauge student states: online textbooks. In online textbooks, student interactions with learning objects generate continuous log data, and tell us what they are doing as they work through the textbook, information we cannot get with hard copy textbooks (Stigler et al., 2020). At the same time, online textbooks are commonplace in STEM education, and the data streams used in the current paper are basic clickstream data which can be readily collected from most online contexts.

The current dataset comes from an introductory data science and statistics course offered through the psychology department at a large US research university. Our data comes from 233 students who stayed on the course for the entirety of the 10-week quarter. The online textbook was a central component in the course. Students learned the material by working through assigned chapters in the textbook every week before coming to lectures. Synchronous class time was spent discussing new examples or common misconceptions.

The online textbook combines reading material, graphics, videos, R coding exercises, and formative assessment questions intended for a college-level introductory statistics and data science course (see Stigler et al., 2020, for further detail). Figure 1 shows a sample section from the textbook. While the textbook was adopted and used across several universities and high schools in the U.S., the current paper focuses on data generated from one particular class for which the research team has more intimate knowledge on.

Figure 1

Sample section from the textbook



Data processing and analysis

The different types of log data generated by the textbook were combined into one stream of data and sorted by time. Week-level metrics were created to abstract student behavior for the week leading up to the homework submission date, which occurred every Thursday for this class. In calculating time spent on an item, we assume that the time between event log A and event log B is spent working on B. For instance, if a quiz was answered in 00:00:00, and a programming exercise was submitted in 00:05:00, we assume the five minutes were used to work on the programming exercise. In this process, we make the generalization that a student is only working on a task directly before creating a log for that particular task, which may not always be true, such as when a student consults previous chapters to solve a question. In addition, if no activity is logged for over 30 minutes, we assume that the activity directly prior is the last activity of a single visit. This is done to prevent the inclusion of times where students had in fact moved away from the screen.

In our analysis, high and low achievement groups are divided by the mean of four graded quizzes administered throughout the semester, outside of the textbook. Quizzes are thought to be a more proximal measure of learning compared to final grade, which had additional operations such as dropping lowest scores, and was subject to some inflation due to emergency measures prompted by the COVID pandemic. We remove the fifth, optional quiz from the data, as most students did not take it. We exclude the last two weeks' process data from analysis as well, given this lack of ground truth and considering the fact that the textbook was used in qualitatively different ways in the last weeks, i.e., reviewing past chapters instead of covering new material.

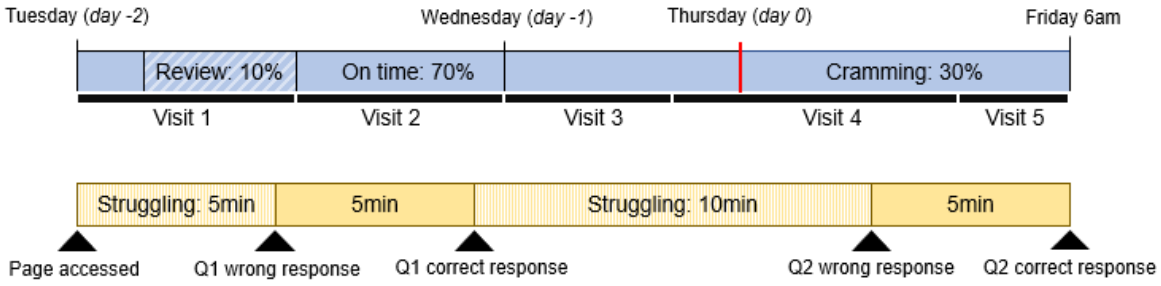
To understand student behavior patterns in the course, two types of metrics were created: proxies of effort, and that of stress and struggling during a week. Time spent is a common, intuitive proxy of the amount of effort students spend (Bodily & Verbert, 2017), while we investigate the difference in visit patterns with the number of days, visits, the ratio of time spent 'cramming', i.e., on the due date and up till 6am the following day, and the earliest day (after Friday 6am) a student visited the textbook for the week. We expect to see more evidence of distributed, in-time work by high-achieving students, and aim to see if such indices could serve as grounds for effort-oriented feedback. Additionally, it is known that the way students review previous or future material is an indicator of their engagement and impacts mastery (Zhao et al., 2014), and we so include metrics on previewing and reviewing activity, investigating whether these indicators differ by achievement group.

For stress and struggling, the number of attempts made prior to reaching mastery has been previously studied as an indicator of struggling and wheel-spinning (i.e., engaging in unproductive work, see Adjei, Baker, & Bahel, 2021 for a review). To number of attempts, we add the *time* spent struggling, as this is expected to reflect the affective state of the student in new ways than count. For instance, even if a student only has two incorrect attempts, if they spent thirty minutes trying to solve this question, this could indicate higher stress than indicated by count. Similar to what has been done in wheel-spinning research, we consider the prevalence of cases where students did not reach mastery, proxied by the proportion of questions students ultimately did not get correct. Affectively, this is considered to be an act of a student 'giving up' on a question and an indicator of stress and struggling, because students were given immediate automatic feedback on an attempt and allowed to make an unlimited number of submissions.

In sum, the resulting effort metrics are the amount of effort made measured by total time spent (*time)*, number of days visited (*num_days*), and number of visits, with consecutive visits divided by more than 30 minutes of no log activity (*num_visit*). The type of effort made is estimated through the proportion of total time spent 'cramming', i.e., on the due date and up till 6am the following day (*ratio_cramming*), the earliest date the student visited the textbook in a given week (*num_earliest*), the ratio of time spent reviewing previous chapters (*ratio_prevchapters*) or previewing upcoming chapters (*ratio_futchapters*), as well as the total number of times a student went back and forth between chapters during the week (*num_transitions*).

Stress and struggling is approximated by the proportion of total time spent before making submissions that were incorrect (*ratio_struggle*), the maximum and average period of time spent continuously in this 'struggling' state (*maxlen_struggle, len_struggle)*, the proportion of total attempts that was marked as incorrect (*ratio_incorrect*), and the proportion of total questions that the student ultimately did not get correct (*ratio_gaveup*). Figure 2 schematizes how the metrics are created from the clickstream data.

Figure 2
Effort and struggle metrics created from clickstream data



For instance, this fictional student has made 5 visits across 4 days since their earliest visit on Tuesday (quantified as 0-(-2) = 2). They spent 30% of time 'cramming', 10% on previous chapters, 0% on future chapters, and made one chapter transition, during visit 1. In the yellow bar, we see that of their submissions, 50% was incorrect, and 60% of their time was spent struggling (15/25). Their maximum and average length of

struggling are each 10 and 7.5 minutes. They have ultimately solved all questions in this simplified example, so the ratio of questions given up is 0%.

For these metrics, we test for group differences and change over time by regression analyses with an interaction term between week and achievement group. The trends of metrics are also visualized for a qualitative discussion of effort and struggle patterns for low and high achievement groups throughout the semester.

Results and discussion

Our first research question asks whether high and low-achieving students show different patterns of effort and struggling over time. Statistical tests reveal high-achieving and low-achieving students do in fact spend their time outside of class differently, and that this difference tends to become more pronounced over time. The number of days visited, number of visits, and number of days between the earliest visit date and the due date decreased over time, and these effort metrics were on average higher for high-achieving groups. Specifically, high-achieving groups visited on average 0.47 days more (*num_days*, $t(1596) = 3.01$, $p = 0.003$), made 1.41 more visits (*num_visits*, $t(1596) = 2.52$, $p = 0.012$), and visited 0.73 days earlier (*num_earliest*, $t(1596) = 3.11$, $p = 0.002$) than low-achieving students for a week, at the start of the semester.

For the type of effort made, we observed that the ratio of time spent on upcoming chapters stayed relatively stable over time, but was significantly higher for high-achieving students by an average of 3.5 percentage points when controlling for week ($t(1596) = 2.91$, $p = 0.004$). While high-achieving students spent about 4.5% of their time on upcoming chapters, low-achieving students only spent on average around 0.01, or 1% of their time on previewing material. The ratio of time spent on previous chapters, which on average was lower for high-achieving students, decreased over time (effect of time $b = -0.05$, $t(1596) = -11.70$, $p < 0.001$), but the rate of decrease was in fact marginally *lower* for high-achieving students (difference in slope $b = 0.01$, $t(1596) = 1.90$, $p = 0.057$), perhaps indicating that some part of the decrease in reviewing behavior may be attributed to an inability to make the effort, rather than a decreasing need.

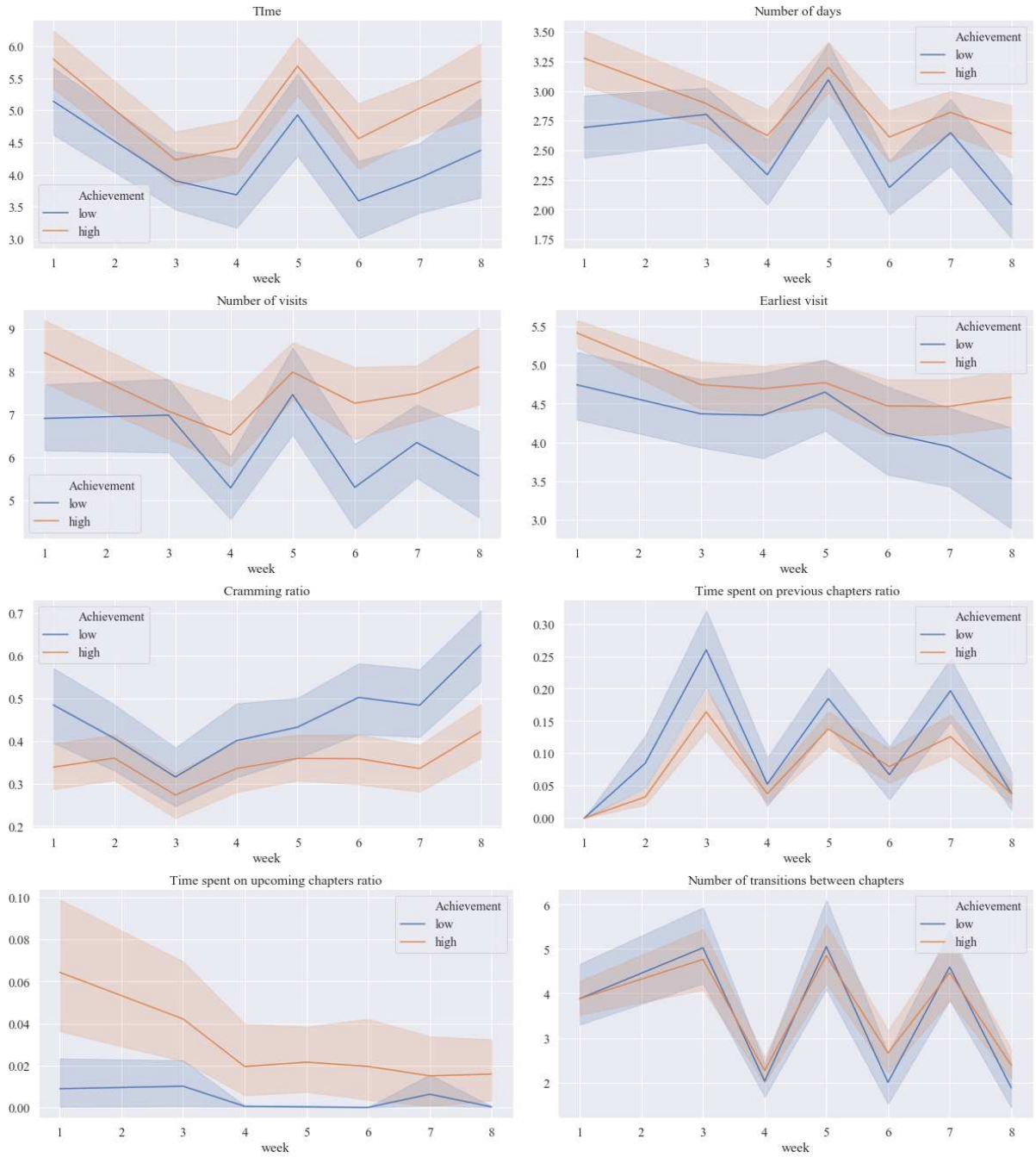
The proportion of questions given up increased over time, and high-achieving learners gave up on about 0.11, or 11.01% of questions at onset (difference in intercept $b = -0.07$, $t(1596) = -4.07$, $p = 0.0001$), compared to low-achieving learners who gave up a much higher proportion of 18.03%. More alarmingly, this gap increased over time (different in slope $b = -0.009$, $t(1596) = -2.39$, $p = 0.017$). A similar pattern was found for the ratio of time spent in a struggling state, with high and low achieving students each spending 12.62% and 17.31% of their time in a struggling state at onset (difference in intercept $b = -0.05$, $t(1596) = -3.07$, $p = 0.002$), and this ratio growing less quickly for high-achieving students (different in slope $b = -0.007$, $t(1596) = -1.99$, $p = 0.047$). The average length of time spent struggling did decrease slowly over time, but the rate of decrease was again significantly higher for high-achieving students (difference in slope $b = -11.47$, $t(1596) = -2.38$, $p = 0.017$). Other metrics (*time, ratio_cramming, ratio_incorrect, maxlen_struggle, num_transitions*) did not show a statistically significant difference between groups in neither the intercept nor the rate of change in our analyses.

In sum, the results show that high-achieving students spend more time on the textbook every week, visiting earlier in the week and splitting up their work across more days and more visits. They are also spending less time ‘stuck’ in a struggling state compared to low-achieving students, a gap that becomes larger over time. While setbacks and confusion are a natural and sometimes beneficial part of learning, extended periods of time struggling alone are less likely to be productive. Indeed, research shows that extended periods of frustration during learning are associated with boredom and disengagement (Liu et al., 2013; Kai et al., 2018). This vicious circle may be one reason why low-achieving students give up on more and more questions as the semester progresses, at a faster rate than high-achieving students. Another reason might be that low-achieving students have a harder time managing workload as the semester progresses, corroborated by the fact that high-achieving students can afford to spend more time previewing upcoming material, while low-achieving students spend more time having to review past chapters.

Next, we qualitatively explore the difference between groups with visualizations of week-level metrics. Figures 3 and 4 illustrate how effort and struggle metrics changed over time for high and low achieving learners, each represented by orange (high) and blue (low) lines.

Figure 3

Week-level effort metrics of high and low achieving students, with shaded 95% confidence intervals

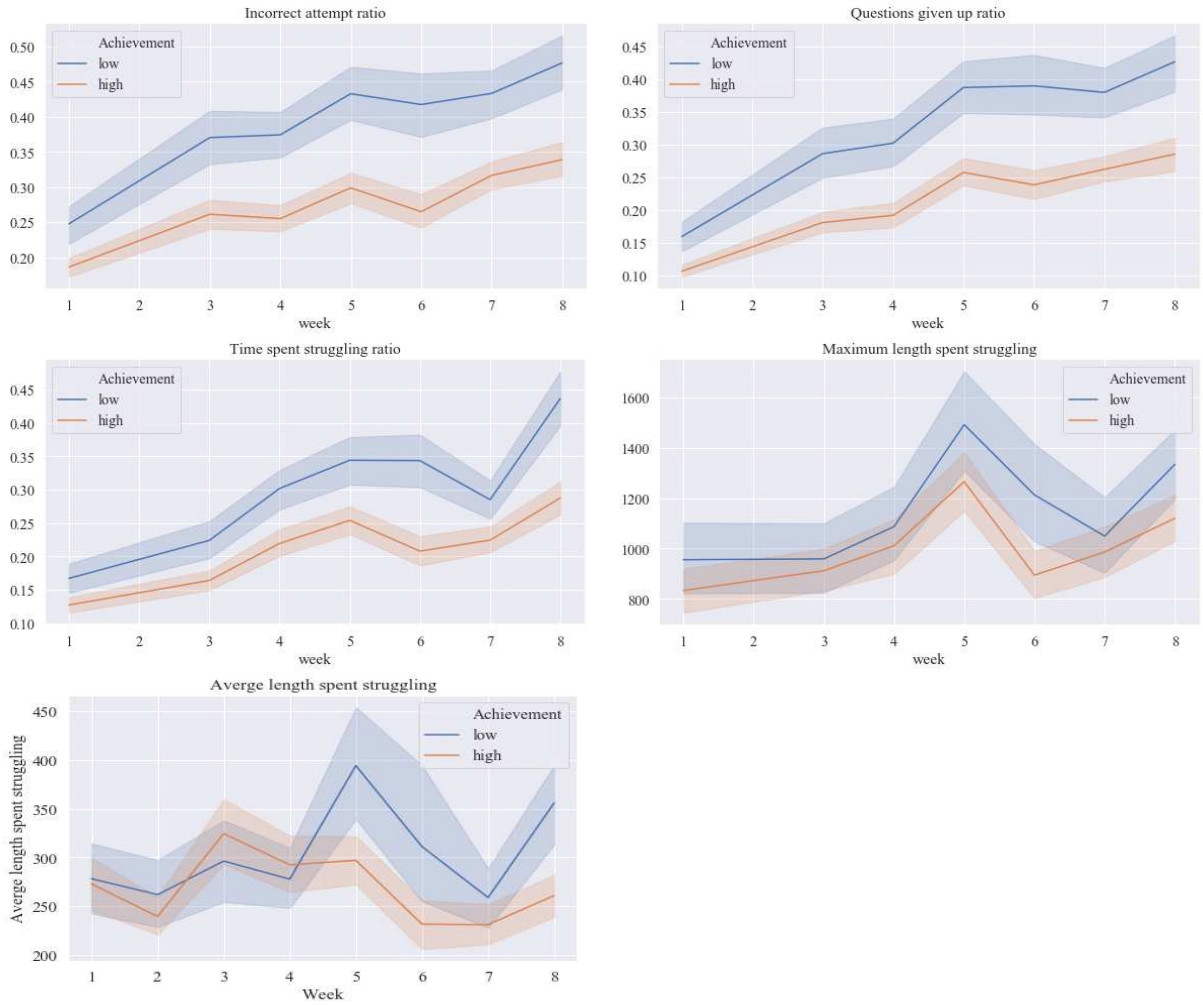


We are able to observe a persisting difference between groups in most effort metrics. For the number of days and visits, the high-achieving group seem to visit relatively earlier and more at onset then again at the end of the semester (a stronger 'w' shape). One possible explanation is that this is a strategic use of time linked to higher achievement. That is, students are allocating more time at the start of the semester to assess workload, adjusting time based on this assessment, then putting in more time as the final exam draws nearer. The cramming ratio of high and low achieving students shows signs of separation in the first week, again indicating a deviance in early time management strategies. Later in the semester, the cramming ratio of low-achieving students increases rapidly together with total time spent, painting a picture of students increasingly struggling to keep afloat of work despite putting in more time.

There is less visible separation in the ratio of time spent on previous chapters, or for the number of transitions made between chapters. Low achieving groups seem to visit previous chapters at a slightly higher rate, and this gap is pronounced in the more difficult chapters, i.e., in weeks 3, 5, and 7, where students made more incorrect attempts as per the uppermost left graph in figure 4. On the other hand, high achieving groups

consistently spend more of their time previewing upcoming chapters, suggesting they have more breathing space in cognitive resources or time.

Figure 4
Week-level metrics of student stress and struggling, with shaded 95% confidence intervals



For stress and struggling metrics, the time spent struggling, the ratio of questions never answered correctly, and the ratio of incorrect attempts seem to increase with time for both groups, but more so for low-achieving groups. By week 8, low-achieving groups are giving up on nearly half of the questions in a chapter (42.61%), and getting nearly every other attempt incorrect (47.66%). Graphing the average and maximum length of time students spend struggling shows that these metrics actually stay stable until mid-semester, until students hit a particularly difficult chapter in week 5.

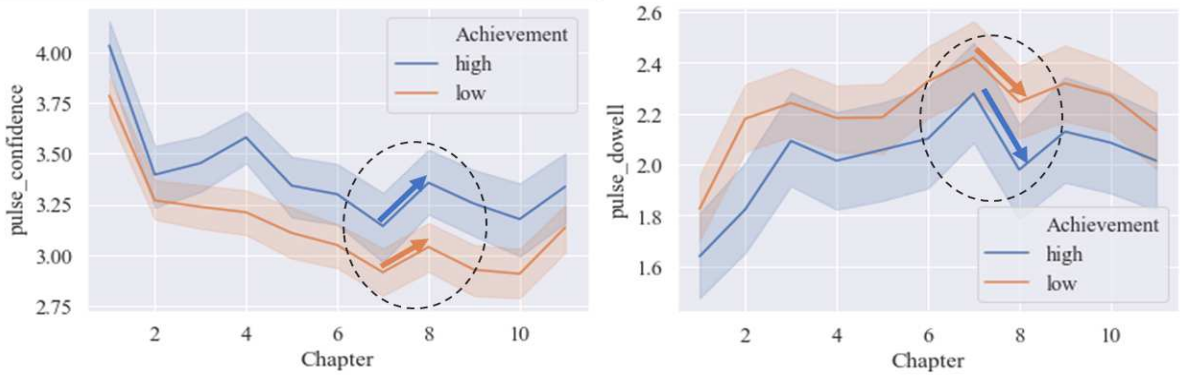
This brings us to our second research question. Struggle-related metrics seem to provide indications as to when 'watershed' weeks are located; weeks where many students falter, and in the current case, low-achieving students disproportionately so. In week 5, low-achieving students found themselves spending much longer periods of time getting questions repeatedly incorrect. In this week, the average time that a low-achieving student spent getting questions consecutively wrong was around 400 seconds, or six and a half minutes. The longest stretch of time they spent stuck in this state was on average around 1500 seconds, or 25 minutes, which is an extremely long time to spend struggling alone in a confused state.

Accordingly, low-achieving students show a momentary drop in the number of attempts, and the proportion of time spent struggling *after* this difficult week, most visible in the 'valley' that is formed across weeks 6 and 7 for the average and maximum length of time spent struggling. High achieving students display a markedly less pronounced drop. This could be indicative of exhaustion or demotivation after a difficult week, more strongly felt for low-achieving students. Supporting this hypothesis, low-achieving students seem to report

a slower recovery of decreased confidence and increased cost perceptions (level of concern on being "unable to put in the time needed to do well in this course.") in their check-up survey after week 5.

Figure 5

Self-reported student confidence and perceived cost across chapters, with critical weeks circled



The right graph in figure 5 highlights how both groups of students report lower confidence at chapter 7, which is the chapter assigned at week 5. The left graph also shows that both groups of students express increased levels of concern on being unable to make time at chapter 7. However, the two graphs show that low achieving students recover less from their drop in confidence and spike in cost concerns afterwards, i.e., a less steep slope between chapter 7 and 8, represented by the orange arrows.

In sum, both quantitative and qualitative analysis of effort and struggle metrics reveal a diverging pattern of behaviors that points to depreciating motivation and academic well-being for low-achieving students. While surveys or scores could also be used to identify these patterns, behavioral observations offer additional contextual information that could help us better support struggling students. For one, the learning material could be updated in an evidence-based way, targeting the parts where students had the most trouble. With item-level information, an instructor could adjust questions where students spent long periods of time struggling. Or, considering the fact that low-achieving students spend more time reviewing previous chapters, a particularly difficult chapter might be scaffolded by adding a summary of relevant concepts from previous chapters.

More importantly, we argue that by implicitly connecting student success to effort rather than ability, and by visualizing student struggles, these metrics have the potential to precipitate a shift towards a more resilience-supportive classroom culture for at-risk students. For instance, being aware of watershed weeks could make it easier for instructors to acknowledge struggles as they occur, which in itself can normalize failure and improve sense of belonging (e.g., Walton & Cohen, 2007, 2011). Indeed, one student we spoke to noted that while the sudden increase in difficulty at week 5 came as a shock, she felt better after learning that her group of friends felt the same way – an experience not every student might have on their own.

Lastly, quantifying and tracking student effort is a step towards a system that can support task-specific, effort attributional feedback; feedback that praises students on their effort and ability to learn from their mistakes, and gives feedforward advice on how to guide their efforts (e.g., 'If you're stuck, try reviewing concept X'). Attributing success to internal, transient factors such as effort is known to support long-term motivation and adaptive beliefs about learning (Schunk, 1983; Cauley & McMillan, 2010). This may be a particularly impactful shift in the introductory STEM classroom, where performance-oriented, competitive classroom culture is known to have enormous negative impact on the achievement of underrepresented groups (Canning et al., 2019) and prime new generations of STEM professionals to this adversarial mindset.

Conclusion

The current paper proposes a set of metrics for gauging student effort and struggles, designed to aid the understanding of student experiences in an introductory STEM course. Metrics on the amount and type of effort, as well as the amount of struggling a student went through in a particular week were able to illuminate the different trajectories that high and low-achieving students go through. In general, high achieving students spent more time and strategic effort learning the material, and spent less time struggling, while low achieving students tended to be less flexible in their efforts, and spent long periods of time stuck in a confused state when studying outside of class. We also observed that this difference in behavior was amplified over time, particularly so in a watershed week where the length of time spent struggling shot up only for low-achieving learners.

We believe that a refined, expanded set of behavioral metrics on student effort and struggles could help adjust instruction and shift classroom culture to be more supportive of struggling learners. For teaching, knowing how student behaviors change could inform how the learning material might be improved, or help instructors acknowledge struggles in a timely way. For learning, metrics could yield knowledge about what adaptive learning patterns look like at both lower levels (e.g., within one visit), and higher levels (e.g., across weeks), forming the basis of task-specific, effort-oriented feedback.

In future studies, we hope to refine our metrics to discover more nuanced student behavior patterns such as different reasons for looking back and forth between chapters, different types of debugging strategies used on problems, or methods of discriminating states of productive versus unproductive struggling. We also plan to expand our analysis using data from different iterations of this and other courses to find patterns that are generalizable across contexts. Ultimately, our hope is to provide a foundation for classroom practices and automated feedback systems that can successfully motivate the struggling learner.

References

Adjei, S. A., Baker, R. S., & Bahel, V. (2021, June). Seven-year longitudinal implications of wheel spinning and productive persistence. In *International Conference on Artificial Intelligence in Education* (pp. 16-28). Springer, Cham.

Belser, C. T., Shillingford, M., Daire, A. P., Prescod, D. J., & Dagley, M. A. (2018). Factors Influencing Undergraduate Student Retention in STEM Majors: Career Development, Math Ability, and Demographics. *Professional Counselor, 8*(3), 262-276.

Bodily, R., & Verbert, K. (2017). Review of research on student-facing learning analytics dashboards and educational recommender systems. *IEEE Transactions on Learning Technologies*, *10*(4), 405-418.

Burke, R. J., & Mattis, M. C. (Eds.). (2007). *Women and minorities in science, technology, engineering, and mathematics: Upping the numbers*. Edward Elgar Publishing.

Cauley, K. M., & McMillan, J. H. (2010). Formative assessment techniques to support student motivation and achievement. The clearing house: A journal of educational strategies, issues and ideas, 83(1), 1-6.

Canning, E. A., Muenks, K., Green, D. J., & Murphy, M. C. (2019). STEM faculty who believe ability is fixed have larger racial achievement gaps and inspire less student motivation in their classes. *Science advances, 5*(2), eaau4734.

Essa, A., & Ayad, H. (2012, April). Student success system: risk analytics and data visualization using ensembles of predictive models. *In Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 158-161).

Gillani, N., Eynon, R., Chiabaut, C., & Finkel, K. (2021). Unpacking the "Black Box" of AI in K12 education. *Educational Technology & Society, Forthcoming*.

Kai, S., Almeda, M. V., Baker, R. S., Heffernan, C., & Heffernan, N. (2018). Decision tree modeling of wheel-spinning and productive persistence in skill builders. *Journal of Educational Data Mining, 10*(1), 36-71.

Kitto, K., & Knight, S. (2019). Practical ethics for building learning analytics. *British Journal of Educational Technology, 50*(6), 2855-2870.

Liu, Z., Pataranutaporn, V., Ocumpaugh, J., & Baker, R. (2013, July). Sequences of frustration and confusion, and learning. In *Educational data mining 2013*.

Schunk, D. H. (1983). Ability versus effort attributional feedback: Differential effects on self-efficacy and achievement. *Journal of educational psychology, 75*(6), 848.

Stigler, J.W., Son, J.Y., Givvin, K.B., Blake, A.B., Fries, L., Shaw, S.T. & Tucker, M.C. (2020). The Better Book approach for education research and development. *Teachers College Record, 122*(9), 1-32.

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.

Walton, G. M., & Cohen, G. L. (2011). A brief social-belonging intervention improves academic and health outcomes of minority students. *Science, 331*(6023), 1447-1451.

Zhao, N., Wardeska, J. G., McGuire, S. Y., & Cook, E. (2014). Metacognition: An effective tool to promote success in college science learning. *Journal of College Science Teaching*, *43*(4), 48-54.