Investigating Relations between Self-Regulated Reading Behaviors and Science Question Difficulty

Effat Farhana
North Carolina State University
efarhan@ncsu.edu

Teomara Rutherford
University of Delaware
teomara@udel.edu

Collin F. Lynch
North Carolina State University
cflynch@ncsu.edu

ABSTRACT

Reading to learn is a quintessentially self-regulated activity. In order to provide effective support for this activity it is necessary for us to understand how students adapt their self-regulation behaviors within disciplinary reading environments. In this paper, we utilize student response data from a digital literacy platform to examine the association of students' behaviors with the difficulty of questions embedded in science texts. We analyzed 131 distinct physical science questions used in 641 middle school classes within Actively Learn, a digital reading platform. We investigated the association of question difficulty and students' behaviors, including reading, annotating, highlighting, and vocabulary lookups. Our findings show that students found multiple choice questions with multiple correct answers hard to answer and exhibited more reading behaviors when attempting them. Short answer questions appeared to be easier; students engaged in more annotation, highlighting vocabulary lookups when attempting easy short-answer questions compared to difficult multiple-choice questions.

Keywords

Question Difficulty, Student Behavior, Self-Regulated Learning

1. INTRODUCTION

Reading to learn, as students do when engaging with disciplinary texts [35], is a quintessentially self-regulated activity [26]. When presented with a block of text, students can approach it by reading end to end, make notes as they go or not. They can also skipp around for clues, or even explore in larger chunks. How they choose to do so will be driven by their own study habits [38], as well as the context of the assignment itself.

Students who are trying to answer a set of questions typically read differently than students who are trying to master general material [11,18]. As the questions change, so will their behavior. They will, to paraphrase Karl Llewellyn, *read with new eyes* [23]. In order to effectively support students in reading to learn, it is necessary to understand how students adapt their reading and learning strategies when faced with problems at different perceived levels of difficulty and of different types. Understanding these changes will allow us to model their behaviors, identify successful and unsuccessful approaches, and provide effective interventions as necessary.

Prior researchers have shown that reading scientific texts requires both reading strategies and self-regulated learning (SRL) strategies [14, 25, 47]. As Butler and Cartier emphasized, understanding SRL requires understanding students' learning contexts [9]. The context of learning is nested: geographical, socio-economical, within-school, and within-classroom. At the classroom level, students' engagement in learning is shaped by teacher's instructional approaches and by interactions with the teacher and peers [9].

Our goal in this study is to examine how students may perceive question difficulty at the *class-level*, and how students vary their individual reading and self-regulated learning activities in response to it. The context of our study is Actively Learn (AL) [1], an online reading platform that is used in schools in the United States. For this study, we focus on readings and test items in middle school science domains. We answer the following research questions:

RQ 1. How does students' performance vary with question difficulty?

RQ 2. What SRL strategies do students use before and after each question ?

2a. How did SRL strategies vary with question difficulty?

2b. How did reading vary with question difficulty?

We collected log data from 11,832 middle school physical science students within the AL platform. We extracted reading, annotating, highlighting, and vocabulary lookup events from the log traces and we estimated the difficulty level of questions by class level. We compared our difficulty level with a comparable analysis from item response theory (IRT) [18]. And we

evaluated students' reading and SRL strategy usage with question difficulty.

2. LITERATURE REVIEW

Our research draws on prior work in two primary areas: research on self-regulated learning activities in reading-to-learn situations and research on question difficulty analysis from student performance data.

2.1 SRL and Science Achievement

SRL, as described by Zimmerman, involves four regulatory components during learning: goal setting, self-monitoring, self-evaluating, and using strategies to control progress toward a goal [51]. Learners who are more capable at self-regulation tend to set more challenging goals for their academic achievement than those who are less capable [53]. They use self-monitoring strategies to monitor their time on task and to solve conceptual problems [8]. Self-evaluation, in this context, means being able to judge the outcomes of self-monitoring processes [52]. In the process of self-evaluation, a student changes learning strategies to achieve their learning goals [53]. Prior researchers have provided a range of SRL models, these include Pintrich's SRL framework [31], Zimmerman's cyclic phases model [50] and Winne and Hadwin's model [46]. While they rely on different assumptions, all of them frame learning as an active process wherein learners set goals by understanding topics or domains, regulate their cognition processes, and modify behaviors to achieve goals in light of self-evaluation [47, 31].

SRL strategies are linked to subject domains [48]. Researchers have examined SRL strategy usage and academic performance in science in game-based learning [37, 40], classroom settings [4], and in agent-based learning environments [7]. Francois et al. examined students' SRL usage strategies in an agent-based learning environment for human biology, MetaTutor [7]. They found high performing students both took more notes and made more summaries. Low performing students, by contrast, struggled to find relevant pages to attain their subgoals within the system. Andrzejewski et al. examined an SRL intervention in a 9th grade earth science class [4]. They found SRL intervention strategies had different effects on students with different socioeconomic status. Students from minority groups (non-white or economically disadvantaged) benefited more than those in the majority group (white and middle class). Rutherford examined the role of SRL within a curriculum integrated mathematics game, ST Math, and found that differences in students' SRL monitoring was related to their academic performance [37].

Our goal in this analysis is to evaluate students' SRL usage in middle school science reading. Our work is situated in the interactions between SRL monitoring and control—as students engage with text and with embedded questions, they assess the difficulty of the task they encounter and adjust their behaviors accordingly. We operationalize the SRL activities related to reading strategies students would use during the control phase of SRL as annotating [24], highlighting [45], and vocabulary lookups, as we believe that these features serve as proxies for SRL behaviors, and we have studied their relation to question type in a prior publication [16]. Science texts involve key concept words and vocabulary terms. Students' reading comprehension and motivation has been found to decrease due to introduction of concept words [22]. Vocabulary lookups can help students to understand concepts when they first encounter them. Annotation requires that students comprehend text and frame it in their own words [24]. Highlighting texts involve SRL activities through the use of monitoring information and connecting that information to prior knowledge [45].

2.2 Question Difficulty from Student Data

Understanding the difficulty level of test items has a wide range of applications in educational data mining (EDM); this includes work on the optimal arrangement of curricula [21] and on the design of adaptive tests or personalized learning environments and intelligent tutoring systems (ITS) [30]. Item difficulty can be assessed based upon the design of a question and its classroom context [20], or it can be evaluated empirically based on observed student performance in real contexts [28]. This empirical approach is particularly important for the development of practical adaptive learning and tutorial environments. Although the structure of a question specifies the knowledge required, the *operational difficulty* of a task, that is the difficulty for a given student, is dependent upon the class context, the amount of individual preparation or scaffolding provided, the students' skill level, and whether they are working on it individually, as part of a team, or as a whole class.

Consequently, a number of prior EDM researchers have developed a number of domain and student models which can be used to identify structural relationships between tasks and to assess their difficulty based upon empirical performance. These efforts include: work on q-matrices that map items to required skills and levels (e.g., [5]); learning factors analysis and other student performance models such as Bayesian Knowledge Tracing (BKT) (e.g., [10, 13]); and item response theory (IRT) [19]. Item Response Theory (IRT) is regarded as the "gold standard" of estimating question difficulties from student response data. The simplest version of IRT is the "Rasch Model" [32], which associates a skill or ability to each student and a difficulty level to each question.

Different intelligent tutoring systems (ITS) [44] and other learning environments have utilized student-system interaction logs to estimate question difficulty empirically. Pardos and Heffernan for example, extended the BKT model to handle item difficulty in a mathematics tutoring system, ASSISTment [30]. QuizGuide, an assessment system for Java programming [39], predicts subjective difficulty on questions from predefined weights and student performance. The predefined weights were assigned by domain experts. ELM-ART II, a web based Lisp programming tool [40], uses fixed difficulty and weight for each item. A student's knowledge level is updated based on correct or incorrect attempts on each item and difficulty level. Researchers have further utilized student attempts coupled with IRT to estimate question difficulty [33]. Fouh et al., for example, utilized the total number of attempts and guessing behavior to understand difficult topics in a Data Structure course [17]. Additionally, they compared their approach to IRT.

As in this prior work, we focus on using student-system interaction logs to estimate the operational difficulty of our questions; however, as we are particularly interested in variation across instructors, we analyze our data at the class level.

3. DATASET

In this section we describe the Actively Learn platform [1] and our dataset construction process.

3.1 The Actively Learn (AL) Platform

AL is a digital literacy platform aimed at students in primary and secondary (K-12) education. AL is designed to improve students' reading proficiency. The platform allows teachers to assign reading texts as assignments to class with embedded questions, which may include optional automated feedback. Assignments in the AL platform can range from one page to multiple pages. Questions in AL can be multiple choice (MCQ) and short answer (SA), including free texts and fill in the blanks. Teachers may use predefined reading texts and questions available within AL or introduce their own as assignments. MCQs are automatically graded, whereas SAs are not. AL questions are graded on a scale of zero to four. Figure 1 shows a reading text in the AL interface.

Physical science reading texts in the AL platform are organized following the Next Generation Science Standards (NGSS) guidelines [2]. The NGSS for middle school physical science (PS) has four standards: (i) PS1: Matter and its Interactions, (ii) PS2: Motion and Stability: Forces and Interactions, (iii) PS3: Energy, and (iv) PS4: Waves and their Applications in Technologies for Information Transfer. Students are expected to analyze and interpret data (PS1 standard), plan and carry out investigations (PS2 standard), develop and use models, analyze data (PS3 standard), and use mathematical thinking and demonstrate understanding (PS4 standard) [2].

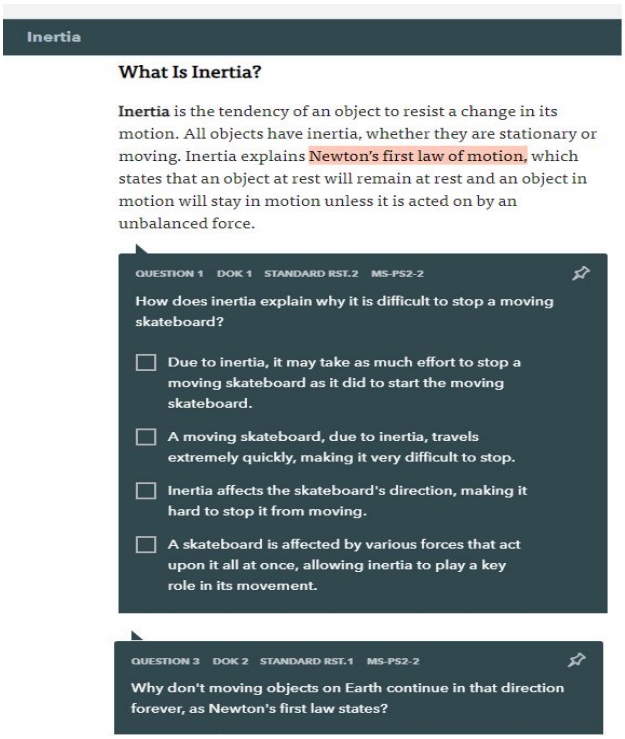

Figure 1. A reading text and embedded questions. Question 1 is an MCQ and question 3 is a SA.

AL's developers state that the platform provides opportunities for teachers and students to deeply engage with text [34]. Students can highlight, annotate, and look up unknown words as they proceed through the readings.

3.2 Dataset Construction

Our current study focuses on middle school science reading assignments in the AL platform. Our dataset includes records of students who completed assignments in 2018. Our dataset includes 17,886 student records across 1,033 classes. After plotting histograms of class sizes, we excluded classes with fewer than 10 or more than 60 students. This left us with 83.45% of students. We also excluded any student enrolled in multiple classes, as we believed these accounts could be for testing purposes. After selecting classes, we filtered the dataset by questions. We selected 131 predefined AL questions used in at least two classes. The final resulting dataset has 11, 832 students and 913 assignments used in 641 classes. We extracted students reading, highlighting, annotating, and vocabulary lookup events from log data trace.

4. METHODOLOGY and RESULTS

In this section we describe our methodology to answer our RQs.

4.1 RQ1: How does students' performance vary with question difficulty?

In our study, a question can be used by different classes. As we do not have access to student demographics and other confounding variables, we opted to aggregate difficulty data at the level of classes. Additionally, we compared our approach with the IRT model. Note that estimating question difficulty is not the goal of our study. We aimed to investigate how students' reading and SRL strategy usage varies with question difficulty. In order to analyze how students respond to different questions, it is necessary to identify suitable metrics to assess question difficulty. First we defined metrics to assess each question difficulty within a class from student interaction data. We analyzed how a question's perceived difficulty varies across classes using our defined metrics. We assessed students' performance on questions categorized by question difficulty. Next, we performed IRT analysis to examine the relationship between question difficulty and student performance. We compared findings between two approaches.

4.1.1 Question Difficulty and Student Performance: Student Interaction Data

We analyzed the students' performance on each question to assess the difficulty of the question. To calculate a student's performance, we took the ratio of max score achieved to number of attempts on a question. Questions in AL are graded on a scale [0-4]. For our assessment, we normalized the students' scores to a range of [0-1]. We defined the performance of a student i on a question q as

$$r_i = \frac{scaled\ maximum\ score\ on\ q}{no.\ of\ attempts\ on\ q} = \frac{maximum\ score\ on\ q/4}{no.\ of\ attempts\ on\ q} \quad (1)$$

Equation (1) computes a student's score of a question on a scale of zero to one, one representing good performance and zero representing poor performance.

We computed difficulty level (*dl)* of a question q as

$$dl = 1 - \frac{\sum_{i=1}^{n} r_i}{n} \quad (2)$$

where n is the number of students in a class who attempted q, and r is the students' performance on q as defined above. A dl ~ 0 value indicates an easy question and dl ~ 1 indicates a difficult one.

To analyze the difficulty of a question q across classes, we computed difficulty ratio of q across classes as follows:

$$\text{Difficulty ratio of question } q = \frac{No.\ of\ classes\ with\ \ dl >= 0.5\ for\ \ q}{No.\ of\ classes\ used\ q\ in\ assignments} \quad (3)$$

We plotted histograms of difficulty ratio for 131 questions. After examining the histograms, we observed more questions with difficulty ratio < 0.2 and fewer questions with difficulty ratio > 0.5. We grouped questions into three categories by their difficulty ratio as shown in Table 1.

We plotted histograms of student performance on each question, *r,* for three categories of questions. Figure 2 presents the histograms (next page).

4.1.2 Question Difficulty and Student Performance: IRT Analysis

The IRT method estimates the probability of a student getting an item correct based upon the item difficulty and the students' ability. We applied the 1-parameter logistic IRT model (1PL) model, also known as the Rasch model. The 1PL model describes test items considering only one parameter, *item difficulty, b.* The 1PL model is a logistic curve, i.e., it evaluates how high the latent ability level needs to be in order to get a 50% chance of getting the item right. Item difficulty is estimated from the student responses.

Table 1: Question category by difficulty ratio (*diff. ratio*)

Question Category	MCQ	SA	Total
Easy (*diff. ratio* < 0.4)	6	75	81
Medium (0.4 <= *diff. ratio* <= 0.6)	5	26	31
Hard (*diff. ratio* > 0.6)	11	8	19

The Rasch model assumes a boolean score for each student response to questions. To apply the 1PL model, we need to map students' responses to 0 or 1 computed from equation (1). We assigned zero if r < 0.5 and 1 otherwise. We fit the 1PL model to 131 questions using the 'ltm' package in R [36].

We plotted per-item characteristic curves (ICC) from the fitted model. The X axis of the ICC represents students' latent ability and the Y axis represents the probability of answering the question correctly. The range of the X axis is [-4, 4], where zero indicates *average* ability. We plotted ICC curves for *Easy, Medium,* and *Hard* questions separately. We also plotted item information curves (IIC) from the fitted model. The IIC curves shows how much information about students' ability an item provides. A difficult item will provide little information about a student with low ability and vice versa for easy items. We plotted IIC curves for *Easy, Medium,* and *Hard* questions separately.

4.1.3 Results for RQ1

From the student interaction results shown in Figure 2. We also notice the number of students receiving zero in *Easy* questions is higher than *Medium* and *Hard* ones.

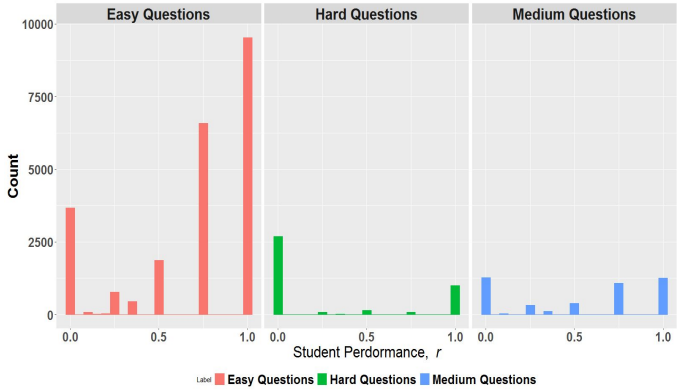

Figure 2. Student performance by question difficulty

In Figure 3 (next page) we show our ICC results for *Easy, Medium,* and *Hard* questions. Each line represents the ICC curve of one question. We observe that the ICC curves for *Easy* questions are mostly on the left side of zero, indicating *Easy* questions required lower ability for correct attempts. Comparing ICC curves of *Easy* and *Hard* questions, we note that *Hard* questions have curves more on the right side of the X axis. The probability of answering a *Hard* question correctly decreases as curves go from left to right.

The IIC curve shows how much information about students' ability a question gives. From Figure 3, we observe *Easy* questions curves provide information about students with average and below average abilities (the peak of curves are mostly on the left side of X = 0. X = 0 refers to average ability). Similarly, IIC curves for *Hard* questions provide information about high ability (the peak of curves are mostly on the right side of X = 0) levels.

4.2 What SRL Strategies Do Students Use Before and After Questions?

In this section we present our methodology and results for RQ2. We calculated SRLs at student-level to understand how students' SRLs varied by question difficulty.

4.2.1 Methodology for RQ2

To investigate the association between students' reading and SRL behavior with question difficulty, first we need to identify student sessions. The AL system does not record student sessions. Therefore, we relied on a data-driven approach to identify sessions as described by Kovanovic et al. [41] and Adithya et al. [3]. AL records timestamps of students' question submission, reading, annotating, highlighting, and vocabulary lookup behaviors. We aggregated timestamps of students' actions into a unified log. We plotted histograms of time intervals between consecutive actions to identify outliers and estimate the last action of any time period [41].

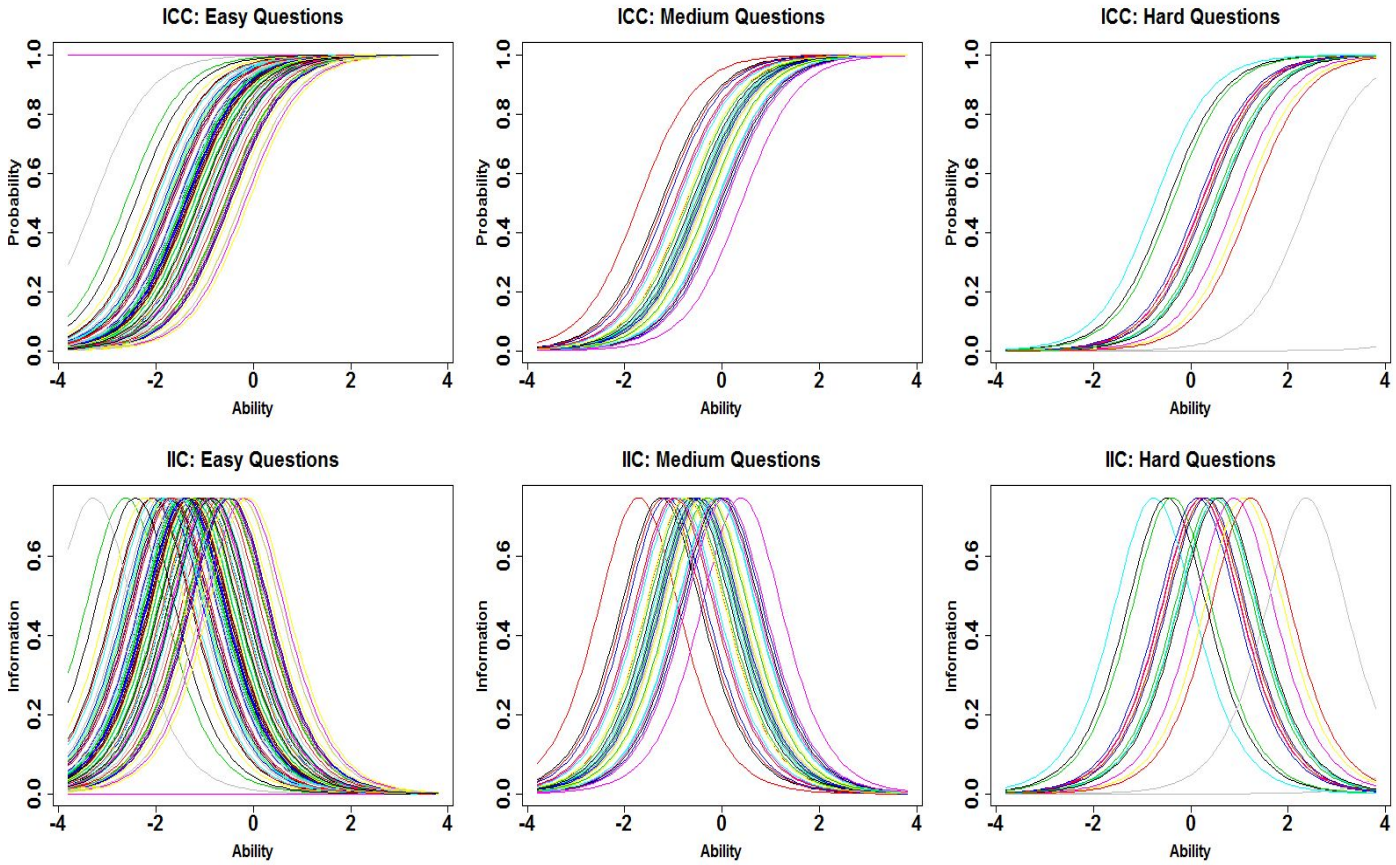

Figure 3. ICC and IIC plots from 1PL model

After conducting this analysis, we selected 30 minutes as a session. Any time interval greater than 30 minutes was marked as the beginning of a new session.

We then split students' actions into sessions. Next, we counted reading and SRL activities prior and after each question submission. We calculated the mean and standard deviation for the four reading and SRL features. To test if there were statistically significant differences in means, we applied the nonparametric Kruskal-Wallis test. In cases with statistically significant differences in mean, we performed a post-hoc Dunn test with Benjamini-Hochberg correction to identify pairwise statistically significant groups, using the R package "dunn.test" [15]. Table 2 presents the mean, standard deviation, and *p* value from Kruskal-Wallis test.

4.2.2 Results for RQ2

In this section we present our results to answer RQ2 and the sub-questions:

2a. How did SRL strategies vary with question difficulty?

2b. How did reading vary with question difficulty?

As Table 2 shows, the mean of all features vary at statistically significant levels across the three categories of questions. Number of reading activities is the highest for the *Hard* questions, followed by *Medium,* and *Easy*. This indicates students had to read more prior to attempting a *Hard* question.

Table 2: Mean with (Standard Deviation), and *p* value from KW = Kruskal-Wallis test for student behavior features on *Easy, Medium,* and *Hard* questions. R = Reading, A = Annotating, H = Highlighting, V = Vocabulary lookups

Feature	*Easy*	*Medium*	*Hard*	KW *p*
R	0.684 (0.71)	0.814 (0.74)	1.27 (0.70)	< 0.001
A	0.335 (0.20)	0.021 (0.17)	0.012 (0.12)	< 0.001
H	0.007 (0.10)	0.004 (0.06)	0.002 (0.05)	< 0.001
V	0.015 (0.13)	0.014 (0.12)	0.009 (0.1)	0.01

It also indicates that they revisited the reading material after attempting a *Hard* question more frequently than they did for *Easy* and *Medium* questions. Annotating, highlighting, and vocabulary lookup counts were higher in *Easy* and *Medium* questions as compared to *Hard* ones.We report the Dunn test and statistically significant pairs for each feature below. We report effect-size (*r*) using a nonparametric test, Cliff's-Delta [12].

For the reading feature (R), we found statistically significant differences among all three pairs *Easy-Hard, Easy-Medium,* and *Medium-Hard.* The *p* values of these pairs were *Easy-Hard* ($p < 0.001$, $r = 0.43$), *Easy-Medium* ($p < 0.001$, $r = 0.10$), *Medium-Hard* ($p<0.001$, $r=0.34$)

When we consider the annotating feature (A), we also found statistically significant differences in means among all three pairs. *Easy-Hard, Easy-Medium* and *Hard-Medium* pairs had ($p < 0.001$, $r = 0.02$), ($p < 0.001$, $r = 0.012$), and ($p = 0.018$, $r = 0.01$), respectively.

And, when considering the highlighting feature (H), we found two pairs differed at statistically significant levels: *Easy -Hard* ($p = 0.004$, $r = 0.004$) and *Easy-Medium* ($p = 0.0209$, $r = 0.003$).

Finally, for the vocabulary lookup (V) feature, we found one pair with a statistically significant difference: *Easy-Hard* ($p = 0.005$, $r = 0.01$).

5. DISCUSSION

We summarize our findings and implications of results below.

In this study we used a data-driven approach on class-level student response data to group questions by difficulty levels. Our difficulty levels are consistent with findings from IRT analysis. ICC curves for *Easy* questions require lower student ability (Figure 3) and vice versa for *Hard* questions.

Table 1 shows 11 MCQ questions belonging to the *Hard* category. We looked into the question texts and observed 10 out of 11 questions required students selecting multiple options, e.g., "Select all that apply." Our analysis from RQ2 indicates students exhibited more reading (R) behavior prior and after answering *Hard* questions compared to *Easy* and *Medium* ones. Thus, our findings indicate that although students can often rule out distractors in MCQs [6], answering such questions is *Hard* when options involve selecting multiple correct answers. Our findings may be helpful for ITS designers. Developers of ITS can facilitate more hints on MCQ questions having multiple correct answers, so that students do not find those *Hard*.

From Table 1, we observed 75 out of 81 *Easy* questions were SAs. Our results for RQ2 indicated that students annotated (A), highlighted (H), and looked up vocabulary (V) more in answering *Easy* questions. We conclude that the format of questions may have contributed to students' SRL usage, even if the difficulty level was classified as *Easy*. Ideally, we would have been able to control question format and student characteristics; secondary data mining allows for large-scale data, but precision of results can be compromised by lack of these details. Nevertheless, we were able to demonstrate that SRL behaviors covary with question difficulty and/or format. It seems likely that as students encountered SA questions, they received metacognitive signals that encouraged their use of SRL behaviors [27] and this resulted in the relatively greater success of these questions. However, we cannot disentangle this from difficulty in our data. Although multiple option MCQs were difficult for students, they may not have triggered metacognitive awareness of the need for SRL behaviors. This is in line with some prior research suggesting less confidence bias in SA questions than in MCQs [29].

6. LIMITATIONS

Our study has two limitations. First, student responses to assignment questions are dependent on the teacher's selection of questions. We do not have responses to all questions for every student. Thus, the latent ability analysis of IRT is limited to student response data. Second, we did not consider the text complexity of the reading article in analyzing question difficulty. Science reading requires analyzing information from texts, diagrams, mathematical equations, and videos [22, 49]. Future research direction can investigate the association of question texts and the reading texts to understand text complexity.

7. CONCLUSION

In this study we investigated associations of students' reading and SRL behavior with question difficulty in middle school science reading. We analyzed question difficulty at the class level and compared our analysis method with IRT. Our results show that MCQ with multiple correct options are generally harder for students in our middle-school set. And we show that when faced with such hard questions, irrespective of their type, students engage in more reading activities but not the other SRL actions we measured. *Easy questions, by contrast, were more commonly* SAs than MCQs. Students spent more time annotating, highlighting, and looking up vocabulary terms in *Easy* questions. This may reflect that the easy questions in our dataset are more focused on rote memorization or on localizing responsive passages in the larger text than on concept synthesis or summarization, or, alternately, SA questions may prompt students to engage in SRL behaviors that MCQs do not. Due to the confounding of difficulty and format type, we were unable to disentangle these reasons. We hope our work opens up further opportunities for researchers and ITS developers to explore student interaction with question difficulty.

8. REFERENCES

[1] Actively Learn. https://www.activelylearn.com/

[2] Next Generation Science Standards.https://www.nextgenscience.org/.

[3] S. Adithya, N. Gitinabard, C. F. Lynch, T. Barnes, and S. Heckman. Predicting student performance based on online study habits: A study of blended courses. In *International Conference on Educational Data Mining*, 2018.

[4] C. E. Andrzejewski, H. A. Davis, P. S. Bruening, and R.R. Poirier. Can a self-regulated strategy intervention close the achievement gap? exploring a classroom-based intervention in 9th grade earth science. *Learning and Individual Differences,* 49:85-99, 2016.

[5] T. Barnes. The q-matrix method: Mining student response data for knowledge. In *American Association for Artificial Intelligence 2005 Educational Data Mining Workshop*, 2005.

[6] L. B. Bliss. A test of lord's assumption regarding examinee guessing behavior on multiple-choice tests using elementary school students. *Journal of Educational Measurement*, pages 147 -153, 1980.

[7] F. Bouchet, R. Azevedo, J. S. Kinnebrew, and G. Biswas. Identifying students' characteristic learning

behaviors in an intelligent tutoring system fostering self-regulated learning. *International Educational Data Mining Society*, 2012.

[8] T. Bouffard-Bouchard, S. Parent, and S. Larivee. Influence of self-efficacy on self-regulation and performance among junior and senior high-school age students. *International Journal of Behavioral Development*, 14(2):153-164, 1991.

[9] D. L. Butler and S. C. Cartier. 2005. Multiple Complementary Methods for Understanding Self-Regulated Learning as Situated in Context. In *Annual Meetings of the American Educational Research Association, Montreal, QC (2005)*

[10] H. Cen, K. Koedinger, and B. Junker. Learning factors analysis-a general method for cognitive model evaluation and improvement. In *International Conference on Intelligent Tutoring Systems*, 2006.

[11] R. Cerdan, R. Gilabert, and E. Vidal-Abarca. Selecting information to answer questions: Strategic individual differences when searching texts. *Learning and Individual Differences*, 21(2):201- 205, 2011.

[12] N. Cliff. Dominance statistics: Ordinal analyses to answer ordinal ques-tions. *Psychological Bulletin.* 114(3):494–509, 1993

[13] A. T. Corbett and J. R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-adapted Interaction*, 4(4):253-278, 1994.

[14] J. Cromley and R. Azevedo. Testing and refining the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology,* 99(2):311-325, 20

[15] A. Dinno. Package dunn.test. https://cran.r-project.org/web/packages/dunn.test/dunn.test.pdf, 2017.

[16] E. Farhana, T. Rutherford, and C.F. Lynch. Associations Between Self-Regulated Learning Strategies and Science Assignment Score in a Digital Literacy Platform. In *Proceedings of the International Conference of the Learning Sciences*, 2020 (In Press).

[17] E. Fouh, M. Farghally, S. Hamouda, K. H. Koh, and C.A. Sha er. Investigating di cult topics in a data structures course using item response theory and logged data analysis. *International Educational Data Mining Society*, 2016.

[18] J. T. Guthrie and A. Wig eld. How motivation fits into a science of reading. *Scientific Studies of Reading*, 3(3):199-205, 1999.

[19] R. K. Hambleton, H. Swaminathan, and H. J. Rogers. Fundamentals of item response theory. Sage, 1991

[20] Y. Hosoda and D. Aline. Two preferences in question-answer sequences in language classroom context. *Classroom Discourse,* 4(1):63-88, 2013.

[21] T. Hsieh and T. Wang. A mining-based approach on discovering courses pattern for constructing suitable learning path. *Expert Systems with Applications*, 37(6):4156-4167, 2010.

[22] Y.-S. Hsu, M.-H. Yen, W.-H. Chang, C.-Y. Wang, and S. Chen. Content analysis of 1998-2012 empirical studies in science reading using a self-regulated learning lens. *International Journal of Science and Mathematics Education*, 14(1):1-27, 2016.

[23] K. N. Llewellyn. *The Bramble Bush: On Our Law and its Study*. Oceana Publications, New York, 1960.

[24] T. Makany, J. Kemp, and I. E. Dror. Optimising the use of note-taking as an external cognitive aid for increasing learning. *British Journal of Educational Technology*, 40(4):619-635, 2009.

[25] D. McNamara. *Reading Comprehension Strategies: Theories, interventions, and technologies.* Psychology Press., 2007.

[26] T. Michalsky. Integrating skills and wills instruction in self-regulated science text reading for secondary students. *International Journal of Science Education,* 35(11):1846-1873, 2013.

[27] H. F. O'Neil Jr and R. S. Brown. Differential effects of question formats in math assessment on metacognition and affect. *Applied measurement in Education,*11(4):331-351, 1998.

[28] U. Pado. Question Difficulty- How to estimate without norming, how to use for automated grading. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1-10, 2017.

[29] G. Pallier, R. Wilkinson, V. Danthiir, S. Kleitman, G. Knezevic, L. Stankov, and R. D. Roberts. The role of individual differences in the accuracy of confidence judgments. *The Journal of General Psychology*, 129(3):257-299, 2002.

[30] Z. A. Pardos and N. T. Heffernan. Kt-idem: Introducing item di culty to the knowledge tracing model. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 243-254. Springer, 2011.

[31] P. R. Pintrich. The role of goal orientation in self-regulated learning. In *Handbook of Self-regulation*, pages 451-502. Elsevier, 2000.

[32] G. Rasch. Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests. 1960.

[33] G. A. Ravi and S. Sosnovsky. Exercise difficulty calibration based on student log mining. In *Proceedings of DAILE, 2013*.

[34] Reading_AL.https://www.activelylearn.com/post/infographic-close-reading-strategies-with-actively-learn

[35] J. S. Richardson, R. F. Morgan, and C. Fleener. *Reading to Learn in the Content areas.* Cengage Learning., 2012.

[36] D. Rizopoulos. ltm: An r package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, 17(5):1 - 25, 2006.

[37] T. Rutherford. Within and between person associations of calibration and achievement. *Contemporary Educational Psychology,* 49:226-237, 2017.

[38] L. Shen et al. Computer technology and college students' reading habits. *Chia-Nan Annual Bulletin*, 32:559-572, 2006.

[39] S. Sosnovsky, P. Brusilovsky, D. H. Lee, V. Zadorozhny, and X. Zhou. Re-assessing the value of adaptive navigation

support in e-learning context. In *International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, pages 193-203. Springer, 2008.

[40] M. Taub, R. Azevedo, A. E. Bradbury, G. C. Millar, and J. Lester. Using sequence mining to reveal the efficiency in scientific reasoning during stem learning with a game-based learning environment. *Learning and Instruction*, 54:93-103, 2018.

[41] K. Vitomir, D. Gasevic, S. Dawson, S. Joksimovic, R.S. Baker, and M. Hatala. Penetrating the black box of time-on-task estimation. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, pages 184 -193, 2015.

[42] G. Weber and P. Brusilovsky. Elm-art: An adaptive versatile system for web-based instruction. *International Journal of Artificial Intelligence in Education*, 12:351-384, 2001.

[43] C. E. Weinstein, J. Husman, and D. R. Dierking. Self-regulation interventions with a focus on learning strategies. In *Handbook of Self-regulation,* pages 727-747. Elsevier, 2000.

[44] E. Wenger. *Artificial Intelligence and Tutoring Systems: Computational and Cognitive Approaches to the Communication of Knowledge.* Morgan Kaufmann, 2014.

[45] P. Winne, J. Nesbit, I. Ram, Z. Marzouk, S. D. Vytasek, J, and J. Stewart. Tracing metacognition by highlighting and tagging to predict recall and transfer. *AERA Online Paper Repository.* 2017.

[46] P. H. Winne and A. F. Hadwin. Studying as self-regulated learning.. The educational psychology series. *Metacognition in Educational Theory and Practice*, 1998.

[47] P. H. Winne and A. F. Hadwin. The weave of motivation and self-regulated learning. *Motivation and Self-regulated Learning: Theory, Research, and Application.*, 2008.

[48] P. H. Winne and N. E. Perry. Measuring self-regulated learning. In *Handbook of Self-regulation*, pages 531-566. Elsevier, 2000.

[49] M.-H. Yen, C.-Y. Wang, W.-H. Chang, S. Chen, Y.-S. Hsu, and T.-C. Liu. Assessing metacognitive components in self-regulated reading of science texts in e-based environments. *International Journal of Science and Mathematics Education.,* 16(5):797-816, 2018.

[50] B. J. Zimmerman. Attaining self-regulation: A social cognitive perspective. In *Handbook of Self-regulation*, pages 13-39. Elsevier, 2000.

[51] B. J. Zimmerman. Self-efficacy: An essential motive to learn. *Contemporary Educational Psychology*, 25(1):82-91, 2000.

[52] B. J. Zimmerman and A. Bandura. Impact of self-regulatory in uences on writing course attainment. *American Educational Research Journal*, 31(4):845{862, 1994.

[53] B. J. Zimmerman, A. Bandura, and M. Martinez-Pons. Self-motivation for academic attainment: The role of self-efficacy beliefs and personal goal setting. *American Educational Research Journal,* 29(3):663-676, 1992