Quantitative and qualitative analysis of student textbook summary writing

Dedra Demaree, Saalih Allie*, Michael Low, and Julian Taylor*
Department of Physics, Oregon State University, Corvallis, Oregon, 97330
** Department of Physics, University of Cape Town, Cape Town, South Africa*

Abstract. The majority of "special access" students at the University of Cape Town are second language English speakers for whom reading the physics textbook is daunting. As a strategy to encourage meaningful engagement with the text, students wrote textbook summaries due the day material was covered in class. The summaries were returned, and they could bring them or re-write them for use during their examinations. A framework was developed to analyze the summaries based on Waywood, defining three cognitive levels seen in mathematics journaling: recounting, summarizing, and dialoging. This framework was refined, expanded, and tested. Interviews with students were conducted for their views on summary writing and survey questions were included on their final exams. The study was carried out in the 2007 spring semester of the "Foundation Physics Course," a component of the special access program.

Keywords: Physics Education Research, Writing to Learn, Textbook Engagement
PACS: 01.40.Fk

INTRODUCTION

During the spring semester of 2007, students in the second semester of the "Foundation Physics Course" [1] (part of the special access program at the University of Cape Town) were asked to write summaries to help them engage meaningfully with the textbook. The course contained 113 students; mostly second language English speakers, who tend to find the textbook daunting. Although writing in conjunction with reading the textbook has been used to help disadvantaged students [2], studies within Physics Education Research have shown less benefit from textbook reading and crib-sheet writing than one would expect [3,4]. As part of this study, we qualitatively assessed the benefits of summary writing, and developed a framework for quantitative analysis of summaries.

THE STUDY

Students in the "Foundation" class were assigned to write summaries of the textbook chapters due the day the material was covered in class. The summaries which covered chapters 3-12 of Reese [5] were scanned and returned. Students were allowed to use either this summary or a rewritten version during examinations. The summaries were restricted to one page (front and back) for each chapter. In addition to the summaries themselves, qualitative data about student views on summary writing were collected. After the semester had been completed, the summaries were analyzed in detail for developing a framework and coding scheme. The framework was tested for reliability and validity, and then applied to analyzing the student summaries.

Qualitative Perspectives

Students in the "special access" program are often insecure, have a weak epistemology, and do not read much else besides the lecture notes. One of the aims of the exercise was therefore to make students realize that the textbook could be used as a significant resource that could expand their own understanding of the material, while more broadly pointing the way to reading as an aspect of taking charge of one's own learning. It was felt that writing summaries would focus the students on what they were reading and activate some degree of sense-making of the text. The feature that the summaries could be used during examinations was introduced to provide "authentic worthwhileness" to encourage full participation.

Eleven representative students were chosen for interviews. One question they were asked was whether they felt their summary would be useful if given to a classmate to use on the exam. Students seemed to feel strongly that their summaries would only be useful to students who were at about the same achievement level as themselves. They felt that students who were better in the course would require different details and those that were not doing as well would need more explanation than what they chose to include. We also found a strong distinction on the part of the students as to writing a summary for the purpose

of understanding the text, versus the purpose of having help on the exam. On the final exam, students were asked to rank two questions on a 5 point scale: how useful the summaries were during the test, and how useful they were in preparing for the test. The average was about 3 for usefulness during the test and about 4 for preparation.

It was observed prior to the summary writing that students expressed the overall sentiment that having summaries (or the textbook) available during their test would be very useful. However, after the summary writing process students were almost unanimous in the view that this would not necessarily be helpful unless you really understood the contents of these resources. Another aspect noted in the discussions with the students was that many of them expressed surprise that they had been able to accomplish reading what at first sight had appeared to be an impenetrable thicket. Anecdotally, students reported that the summary writing provided a good focus for engaging with the book, and that it gave them confidence in being able to undertake future readings on their own.

Quantitative Perspectives

As part of an ongoing effort to quantify writing quality, a framework was chosen from which to create a coding scheme. Details on the research methodology can be found in the accompanying paper in this volume [6]. The starting point was the Waywood framework [7] defining cognitive levels of student engagement in mathematics journaling: recounting, summarizing, and dialoging. Recounting means "telling what happened." For the purpose of summary writing, this translates to re-copying things straight from the text. Summarizing is an attempt to put things in your own words, and represents a somewhat higher level of engagement. Dialoguing includes indications of sense-making, such as places where students pose questions about the material.

A detailed look at a large subset of the summaries yielded categories for what types of content students included, as well as what other features could be coded such as global organization. Types of content observed and coded included definitions, equations, derivations, methods, diagrams, and graphs. A method is something such as a generic problem solving technique. We coded these items based on Waywood, and expanded the coding to include details for what would situate each content type into each category. For example, with equations, recount would mean simply copied from the text. Summarized would be an equation stated in words or with each variable defined. A dialogued equation might include details such as how it could be applied.

The Waywood framework made no distinction as to what content students chose to include, only *how* they included it, so a new category was added to the framework. We called the new category "content choices," and used it rate the importance of the content students chose to include. For example, we wanted to distinguish between students who focused on equations that cover all cases versus students who had multiple forms of the same equation each applying only to a specific case.

In addition, we coded global issues such as organization. We looked to see if students differentiated between types of content, for instance labeling examples with the word "example." We noted if the content was grouped in any specific way, or if it followed the order presented in the book. We also looked to see if the students outlined any sort of hierarchy of importance in their summary, by providing headings for main ideas, or underlining, bulleting, or highlighting important ideas.

Our modified framework and coding scheme attempted to capture the content and features of the student summaries as fully as possible. The framework also provided several ways to assess student engagement with the text: via the choices they made, the level of cognitive depth of how they chose to include the content, and the global organization.

Coding and Numerical Data

Each summary was coded sequentially to ensure that all content in the summary was included in the coding. Simultaneously, attention was paid to the degree to which text items, such as a particular definition, were copied directly from the book. A few difficulties were found in applying the coding scheme. Sometimes students put the same equation multiple times in different forms (for example $F = ma$ and $a = F/m$), making it hard to choose if this should be coded as multiple equations, or as an attempt to go beyond simply recopying the equations. It was also hard to decide if some content was a method or dialogue belonging to some other content. If a method was connected to a specific equation it would instead be coded as a dialogued equation.

It was hard to choose if equations were general or specific because in most cases where students included specific ones they were embedded in derivations or examples. The most significant category dropped from the coding scheme was distinguishing the importance of choices students made about definitions. It was too subjective to determine if a certain term being used *should* be defined as this would depend on the mindset of the student regarding the usefulness of the term in the context it was being employed.

For the purposes of analysis, summaries from 15 students were coded. These students were chosen based on the fact that we had a full set of summaries for them, and they represented a mix of achievement levels as measured by their final grade. The coding scheme allowed us to give quantitative scores across several aspects of their summaries. Scores were assigned on the basis of their "content choices," their "content" as coded by the Waywood cognitive levels, and for the "global issues", as defined above.

For the "content" scores, each item counted as being worth 3 points if categorized as "dialogue," 2 if "summarize," and 1 if "recount." The total was divided by the number of items in that category, giving a normalized score for the content in each content category. Scores based on "content choices" and "global issues" were averaged over each subset of those categories. A total score was then given to each summary based evenly on each of the three sub-scores.

Two features of the coding specifically represented writing where students went beyond simply stating the physics facts. That was including methods content (explanations of how to do things), and content that fell in the "dialogue" category. We therefore also created a score based on the number of times students did each of these things, thinking that this might be an alternative way to score the summaries.

Testing the Framework

In applying the framework, the following questions were addressed: (1) Do independent researchers code the summaries reliably using the framework? (2) Is the framework valid as measured by comparing the quantitative characterizations with qualitative rankings made by independent instructors? (3) Are any features more indicative of good quality than others? (4) What is different between pre and post summaries? (5) How can we further refine the framework? The chapter 4 essays were of particular interest since all students chose to re-write those for the final. We called these 2nd drafts the "4t" summaries since they were specifically written for taking to the test.

Validity of the Framework

In order to test the validity of the framework, we asked 6 instructors to independently rank the 4t summaries. The professor of the course ranked them, as did two of the other authors on this paper. 3 experienced instructors and professors not familiar with this project also ranked the summaries. The instructors ranked the students into 3 categories: low, medium and high. An instructor score was created using these rankings. A Pearson correlation test showed significant agreement ($r = 0.535$ at the 0.05 level, 2-tailed) with the instructor "score" and the total score calculated based on content, choices, and global issues, showing validity of the framework.

Agreement with the instructor score was strong but not quite significant for the coding based on the content score ($r = 0.347$) and the number of things in the dialogue category ($r = 0.370$), but there was no correlation ($r = -0.004$ at the 0.990 level) for the number of comments in the methods category. Counter to our intuition that the inclusion of methods content might be an indicator of a good summary, this did not seem to be borne out by observation. The most strongly correlated item ($r = 0.586$ at the 0.05 significance level) was the score based on the global issues, possibly indicating that instructors are likely to take their cue from certain surface features in determining whether a summary is good or not.

Reliability of the Framework

In order to test the reliability of the framework, an independent coder coded the 4t summaries using the framework and coding scheme. He felt that the framework was straightforward to use and was able to account for all the content. The difficulties he experienced were similar to those described above. He found it hard to code equations if there were multiple ones covering the same idea. He also found it hard to choose between "recount" and "summary" if student writing was not quite identical to the textbook.

Comparing the two sets of coding, we found the global features were coded nearly identically. The overall scores as computed by the coding were quite similar between the two coders as well. The main source of difference was between the "recount" and "summary" categories, specifically for content containing definitions. The independent coder was more likely to put equations in the "recount" category than the original coding. There was also discrepancy between coders for the "methods" category, probably the least well defined in the framework, and in need of future revision.

Coding Correlations

Of great interest to us was which aspects of the coding most strongly correlated with the overall summary scores. It seemed that some specific features of the summaries were more indicative of overall good summaries than others. The total score given to the content correlated most strongly with the scores given to the definitions and equations ($r = 0.718$ and $r = 0.703$ respectively, Pearson correlation, at the 0.01 level, 2-tailed). This score also correlated well (but not quite significantly) with derivations ($r = 0.497$) and methods ($r = 0.429$), but not significantly with graphs

(r = -0.130) and diagrams (r = -0.246). In addition, students who had high scores for global issues were very likely to have high scores for derivations (r = 0.759 at the 0.01 level) and equations (r = 0.621 at the 0.05 level). This suggests that a reliable score of the quality of a summary might be achieved with a reduced set of coding items.

Pre versus Post Analysis

To do detailed analysis of pre and post summaries, we looked to see if the coding scheme would measure a difference between the chapter 4 and 4t summaries since they were written pre versus post instruction as well as for different purposes: sense-making versus exams. Overall we didn't see a significant difference in total scores for these two types of summaries. It is interesting that our coding scheme did not see any differences, and indicates that either the scheme needs to be more nuanced, or that although the purposes are different, the outcomes are not very different.

We found that on average there were 25% fewer copied equations in 4t vs. the original chapter 4 summaries (7.5 in 4t vs. 10 in the 1st drafts). There were also less copied ("recount") definitions in the 4t summaries, possibly indicating that they did not include definitions of terms that they had learned well. Though the total number of derivations was the same, there were more coded in the summarize and dialogue category in the 4t summaries, indicating an overall improvement in quality. Although all examples included in the 4t summaries were in the recount category, it seemed the students made better choices about which examples to include in the 4t summaries.

By looking at the choices category, overall it appears that students made better choices about what to include on the 4t summaries than the 1st drafts. It is possible they chose to cut some things out of the 4t summaries because they found that information was no longer necessary (such as simple definitions, or trivial diagrams and examples), but we cannot confirm this without further interviews.

CONCLUSIONS AND COMMENTS

This study had two different purposes. One was to help students engage with the text. The other was to develop a framework and coding scheme to quantitatively analyze student summaries. We found that students were empowered by the process of engaging with the text and were proud they had been able to tackle reading a seemingly intimidating textbook. Students felt that writing the summaries was a useful activity and helped them prepare for the examinations.

When developing the framework we answered the questions posed above. We found (1) that the methods category is not well defined, but most of the coding matched well between two different coders; (2) that the scores from the coding matched well with qualitative rankings given by independent instructors; (3) that the cognitive level of the included equations, definitions, derivations, and methods, along with the global issues were the strongest indicators of the overall summary quality; (4) very few differences between pre and post summaries contrary to indications from student interviews. (5) We propose demarcating the methods category more clearly and refining the choices category to make the coding easier to use for different researchers. In addition, possibly removing the coding of diagrams and graphs would save time without loss of coding quality.

Overall, the framework appears to be valid and reliable, and usable by other researchers, providing a quantitative way to analyze summaries. However, we found that it was difficult to tell why students made the choices they did, as it depended on their level of understanding, the amount of time they put in, and other factors that cannot be gauged without additional surveys or interviews. We feel that the role of audience strongly confounded our ability to score the summaries. Each student is of course writing for an audience of one, themselves, and therefore has a different purpose, a different set of needs, and choose different things as important. In future studies it would be interesting to control for the role of audience by specifying one in the summary writing assignment.

ACKNOWLEDGMENTS

The authors of this paper thank Nam-Hwa Kang, Len Cerny, Elizabeth Gire, and Corinne Manogue for their input and assistance with this study.

REFERENCES

1. S. Allie and A. Buffler, *Am. J. Phys* **66**, 613-624 (1998).
2. C. Kalman, M.W. Aulls, S. Rohar and J. Godley, *Jrnl. Coll. Sci. Teach.* **37**, 74-81 (2008).
3. N. Podolefsky and N. Finkelstein, *Physics Teacher*, **44**,338-342 (2006).
4. K. Harper and S. Doty, *What Makes a Good Crib Sheet?,* AAPT Summer Meeting: Salt Lake City, UT (2005).
5. R.L. Reese, University Physics, Brooks/Cole Publishing Co., Pacific Grove, CA (2000).
6. S. Allie, D. Demaree, J. Taylor, F. Lubben and A. Buffler, "Making Sense of Measurements, Making Sense of the Textbook" in *Physics Education Research Conference Proceedings,* edited by C. Henderson et. al., Edmonton, Canada, 2008.
7. A. Waywood, *For the Learning of Mathematics* **12**, 34-43 (1992).