Content analysis of instructor tools for building a learning community

Carissa Myers,[1] Adrienne Traxler,[1] and A. Gavrin[2]

[1]*Department of Physics, Wright State University, 3640 Colonel Glenn Hwy, Dayton, OH, 45345*
[2]*Department of Physics, Indiana University Purdue University Indianapolis, 402 N. Blackford St., Indianapolis, IN, 46202-3217*

This work presents a content analysis of an online discussion forum accompanying a face-to-face introductory physics course. Content analysis is a quantitative method for analyzing text that uses a coding scheme to gain insight into student discussions. We explore the effects of "anchor" tasks, small weekly activities to help students engage with each other. The goal of this analysis was to examine how the distributions of codes are impacted by anchor versus non-anchor tasks, and different types of anchors. The result of this work was that the coding scheme was able to detect some differences between anchor and non-anchor threads, but further work should be done to observe behaviors that would require a more in-depth analysis of the text. This research is significant for physics education research (PER) because there is little PER using content analysis or studying online talk. This is a step towards identifying patterns in conversations between physics students and the tools that may help them have on topic conversations essential for their learning. Identifying such tools can aid instructors in creating effective online learning environments, and this project introduces "anchor" tasks as instructor tools for building a learning community.

Keywords: content analysis, online learning

I. INTRODUCTION

Content analysis was performed on the conversations from the online discussion forum of an introductory physics course. The coding scheme used in this work was the Transcript Analysis Tool, also referred to as the TAT [1]. We used the TAT to code each sentence with an identifier such as statements, questions (open or closed), and off-topic comments (comments not related to physics). This tool allows for the frequency of sentence types to be studied, which helps identify possible patterns within the code types and frequencies.

The effects of "anchor" tasks on the code distribution were analyzed. An anchor task is a small activity or task posted by the instructor to help students engage in focused forum conversations [2]. At the start of each week, the tasks were posted at the top of the discussion forum. Example tasks included introducing yourself (first week), posing a good physics question as practice for an exam, or sharing an interesting fact about Newton. An anchor thread was each new conversation the students had regarding the task. Non-anchor threads were all the other conversations, where discussions were built on random topics chosen by the students. Non-anchor threads required students to begin a conversation without any supporting tasks given by the professor.

The motivation for this work was a prior social network analysis of three semesters of introductory physics discussion forum data [3]. This analysis found a correlation between central network position and overall course grade for semesters one and three, but not for semester two. This difference could not be explained by the network analysis. Content analysis was introduced to look for sources of the difference within post text. We hypothesize that the position/grade correlation does not exist for semester two because anchor tasks were not used in this semester. To explore the difference between semesters, the first step is demonstrate that content analysis can be applied, and to evaluate transcript analysis tools such as the TAT. For the preliminary evaluation, we used the TAT to code the discussions in semester one.

This research draws connections between how the instructor designed and facilitated the forum and the student conversations that result. Our goal is to explore the overall patterns of conversation using the TAT and to analyze the relationship between anchor threads and code distributions in particular. We will contrast how anchor threads provoke different code distributions in the related posts, and compare anchor with non-anchor threads. This work provides an insight into how students use anchor tasks, and a replication study of the TAT. This is a step forward in validating the TAT. This research is important for physics education because it identifies patterns in online conversations, includes on- and off-topic comments as useful, and identifies tools to help the instructors create effective online learning environments. It also gives a "proof of concept" for adapting content analysis tools from online learning for physics education research.

II. CONTENT ANALYSIS

Content analysis is a quantitative method for analyzing text [4]. In physics education research, content analysis is useful because it establishes a bridge between quantitative and qualitative research. This bridge allows social, behavioral, and cognitive patterns to be studied, which are patterns that are hard to see when using only quantitative methods.The method consists of choosing and applying a coding scheme. A coding scheme is a system of identifiers that are used to tag a unit of analysis such as sentences, thematic units, or whole messages. Coders are trained to use the coding scheme, usually assigning one code per unit of analysis (though some schemes allow multiple codes, see [5]). By categorizing variables of interest and coding large amounts of data, large-scale patterns may be identified for statistical analysis. This analysis may

lead to a qualitative content analysis to further study themes, based on behaviors, perceptions, and social trends.

Content analysis is a highly flexible technique in which many standards are still being developed. Researchers and studies use different units of analysis. Some studies code messages, others use sentences, and others use thematic units (a single thought or idea of varying length [5]). Few studies test the replicability of coding schemes. Both of these can result from how coding schemes are tied to cognitive theory: researchers are looking for different patterns, so they consider different "grain size" and often design new coding schemes instead of reusing prior work.

Coding schemes also vary because content analysis is used differently from field to field. Computer supported collaborative learning (CSCL) studies tend to use content analysis to identify social trends and cognitive levels in discussions. In the few cases where physics education researchers have used content analysis, they tend to code for specific physics constructs. Kortemeyer's work is an example of content analysis being used to identify physics concepts and problem solving techniques in online homework discussions [6]. He identifies themes based on how the students were discussing the topic (*e.g.*, conceptually versus procedurally) and on aspects of a physics problem (*e.g.*, finding the solution versus understanding the physics). The TAT focuses more on identifying the broader types of interactions between students, rather than discipline-specific concepts. In PER, content analysis can help to connect physics constructs with the students' conversations and thoughts.

III. METHODS

A. Course context and data

In this study, content analysis was used to analyze online discussion forums of an introductory, calculus-based physics course. The course was a lecture with active learning, and consisted of 173 students, of which 156 posted in the forum. The instructor was also active in the discussions. There were 936 threads with 2,376 reply comments and a total of 6,396 sentences. If the students participated in the discussion, they could gain up to 5 percent extra credit on the final grade. For more course details, see Traxler *et al.* [3].

B. Coding and inter-rater reliability

Two coding schemes were piloted to decide which best fit the data and the three coders on this project. The first was the TAT [1], which codes sentences by five primary categories: questioning, statements, reflections, scaffolding, and citations. The other scheme was the Community of Inquiry (CoI) framework [5], which is popular in CSCL studies. This model uses thematic units to code for social, cognitive, and teaching presence. After coding a sample of 124 sentences using both schemes, the coders came together as a group to discuss the methods. The CoI proved to be challenging because the coders struggled agreeing on thematic units. The main issue was identifying which pieces of the text were different themes, and there was no consistency using the themes. The TAT provided fewer issues distinguishing between the types of sentences, which lead the coders to choose this model for this work. Unfortunately, a tradeoff between theoretical elegance and coding reliability is common in CSCL content analysis studies [7].

Inter-rater reliability is calculated to measure the agreement between different coders for the same text. If other coders can repeat the process yielding the same results, then the inter-rater reliability strengthens validity for the coding scheme. Common reliability measures focus on whether the coders agree to the exact values assigned for each unit and can flag problematic codes. For this work, percent agreement was chosen since it is easy to calculate and interpret, but it does not account for chance agreement between coders. Since Cohen's kappa corrects for chance agreement while also comparing the similarities between raters, it was also calculated. These two measures are the most commonly reported in quantitative content analysis, and reporting both is recommended in a review by De Wever *et al.* [7]. We also include Fleiss' kappa, an adaptation of Cohen's kappa to more than two coders [8].

The TAT was used to code a sample of data from the first three weeks of the semester, with each sentence marked independently by all coders. Cohen's kappa was calculated to check reliability. Next, coders discussed areas of difficulty such as rhetorical questions and emoticons, the difference between scaffolding and reflection comments, and how to distinguish vertical and horizontal questions. After this conversation, the coders recoded the sample on their own. This increased the agreement between coders, shown in Table I by comparing the Pre-discuss and Post-discuss columns.

TABLE I. Inter-rater reliability for the pilot sample from weeks 0–2 and the second sample. Percent agreement and Cohen's kappa are pairwise; Fleiss' kappa includes all coders.

	Pilot sample		Second sample
Measure	Pre-discuss	Post-discuss	
Percent agreement			
1 and 2	0.46	0.85	0.68
1 and 3	0.42	0.81	0.70
2 and 3	0.69	0.81	0.71
Cohen's Kappa			
1 and 2	0.27	0.74	0.46
1 and 3	0.30	0.66	0.52
2 and 3	0.51	0.63	0.54
Fleiss' Kappa	0.30	0.70	0.50

TABLE II. TAT codes, sentence types, and examples.

1A	Vertical question	• Does anyone understand how to find the tension on one of the pulley problems.
1B	Horizontal question	• How long are you going to study for this test?
		• This might be a long shot, but can we not use electromagnets and the concept of electromagnetic propulsion to put a rocket into space?
2A	Statements that give information	• T and mg are the only forces acting there, and must be responsible for the acceleration.
2B	Direct response to a question	• Just remember to think of each component separately.
3	Self-reflection	• I just learn best by teaching others so I hope we have a good turn out.
4	Scaffolding	• Don't feel bad, you aren't the only one feeling this way about the test.
		• I went home and hunted all weekend and had a blast at a wedding!
5A	Direct references and quotations	• Take a look in the link for the equations and a visual aid.
5B	Citations or web-links	• www.youtube.com/watch?v=kvCnjVSpuv0

To check for inter-rater reliability, around 10 to 20 percent of the data (randomly chosen) should be coded [4]. To reach 10 percent, the second data sample came from weeks 6–11. To randomize the sample, a combination of cluster sampling and systematic random sampling was used [4], by coding the sentences in every fourth thread. Table I shows all three inter-rater reliability measures (percent agreement, Cohen's kappa, and Fleiss's kappa), calculated using the `irr` package in R [8]. The reliability measures in this work all normalize to 1. The values for all the kappas fall in the range between 0.65 and 0.75 for the post-discussion pilot sample. This range is acceptable, though not above 0.8 as recommended by others for high reliability [4, p. 142]. The second sample had lower values, especially for the kappa coefficients. This may be because kappa coefficients are overly conservative for uneven code distributions [4], but is more likely due to non-ideal coding conditions. Coder 2 did the primary work on weeks 6–11, and coders 1 and 3 coded the second sample from this time period after several months had passed.

After increasing inter-rater reliability for the pilot sample, the remaining data was divided into thirds and each portion was coded by one person using the TAT and RQDA package [9]. The TAT tags each sentence with a single identifier, summarized in Table II. Types 1A and 1B are questions. Type 1A are vertical questions, which are closed-ended or have "correct" answers. Type 1B are horizontal questions, which are open-ended questions or questions that have long responses or more than one right answer. Types 2A and 2B are statement types. Type 3 identifies reflection, meaning the person is sharing personal thoughts, opinions, or feelings. Type 4 identifies scaffolding or engaging comments. Engaging comments are meant to initiate conversation and interaction. Type 4 could be identified by searching for greeting, thanking, welcoming, and agreeing comments. Type 5A and Type 5B indicate citations, references, and direct quotations. In this work, there was no value placed more on one code type than another, nor was there value placed more on anchor threads than non-anchor threads. Instead, the focus was placed on whether certain patterns regarding code types (did one code appear more than another, etc.), and anchor and non-anchor threads would appear.

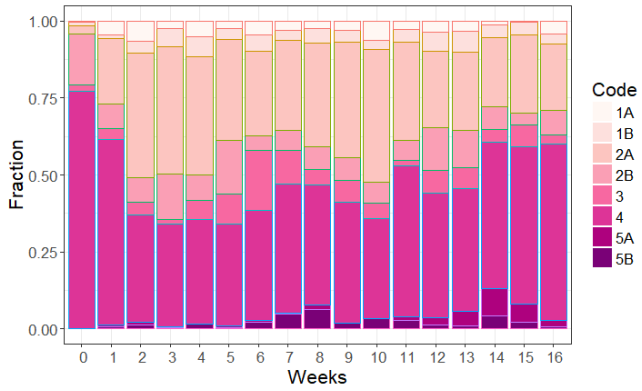

FIG. 1. Percentage of code types for each week in the first semester.

IV. RESULTS

Figure 1 represents the percent of code types for each week. Informational (2A) and scaffolding (4) statements were the most common in all weeks, with the largest number of scaffolding statements in week 0 (posts before the official start of classes). Also, anchor tasks were given each week except 6, 10, and 13–15. As the weeks progressed, the number of anchor threads decreased. At the same time, students started discussing the link between everyday life experiences and physics more.

Figure 2 shows the relative frequency of code distributions within all anchor and non-anchor threads. Anchor threads had a larger frequency of scaffolding statements (TAT category 4) and responses to questions or statements (category 2B). Non-anchor threads had a larger frequency of statements offering information (2A), horizontal and vertical questions (1A and 1B), reflections (3), and citations and weblinks (5A and 5B).

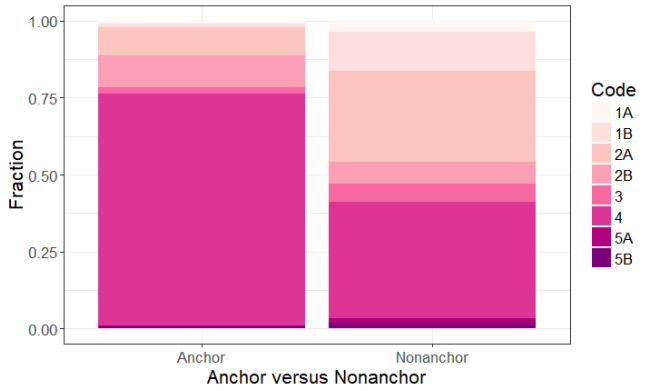

FIG. 2. Frequency of code types in the 171 anchor threads and 765 non-anchor threads.

V. DISCUSSION

The goal of this work was to examine how the code distribution was impacted by the anchor and non-anchor threads, and by the different types of anchor threads. Differences appeared in the code distributions between anchor and non-anchor tasks, with anchor tasks having a larger frequency of scaffolding and questions or responses to statements. This difference may be due to the fact that anchor threads are built to encourage student conversation, which means off-topic statements and responses would make sense. This difference may also be due to the topics of the anchor threads. For example, students may be responding to certain anchor threads more than others because they feel more comfortable or the topic is easier to discuss. Future research would observe whether these differences occur in multiple semesters, reasons for the differences, and what this means in regards to learning physics.

Regarding the different types of anchor tasks, the number of anchor threads dropped as the weeks progressed. The reason for this decrease is not clear, and not identifiable by the TAT. The later anchor tasks required more thought and work for the students; hence, the students may have stopped responding to the anchor tasks because of higher difficulty or late-semester workload in other courses. Alternately, the students may have formed a learning community, and grown comfortable enough to discuss perspectives and ideas without needing the anchor tasks to encourage them. Another observation was that students started posting more about the connections between everyday life experiences and physics. This trend emerged late in the coding process, and the TAT does not distinguish coursework from real-life physics talk. Future research can explore whether this trend appeared in other semesters, and if not, what causes this difference.

The TAT detected differences between anchor and non-anchor threads. It could also detect changes in discussion behavior such as shifts in sentence types. The TAT was straightforward to use based on the unit of analysis. However, some categories were much less reliable than others, and the overall inter-rater reliability was disappointingly low. Anchor threads revealed a major limitation of the TAT for our purpose: it could not observe themes such as students making a connections between physics and everyday life experiences. It could also not detect reasons behind the behaviors, such as whether the anchor threads scaffolded productive conversations. Considering these details, the TAT seems good at observing text on the surface, but missed some behaviors or patterns that would require a more in-depth analysis of the text.

Future research would choose or develop a coding scheme that analyzes features of the anchor threads. This would allow a more in-depth look at their patterns and might link early-semester anchor threads with the later discussion community. The TAT has a place for "off-topic" comments (typically in codes 3 and 4), which are important for building community. To catch nuance such as making real-world connections with class material, a scheme with physics content or finer marking of cognitive level might perform better. After finding a satisfactory scheme, coding and comparing semesters one and two might explain the observed differences in their networks.These results could help instructors fine-tune anchor tasks to more reliably build a productive forum community.

ACKNOWLEDGMENTS

We thank Fenton Clawson for his help in coding the data and Chad Campbell for methodology conversations.

[1] P. J. Fahy, G. Crawford, and M. Ally, The International Review of Research in Open and Distributed Learning **2**, 1 (2001).
[2] M. Guzdial and J. Turns, The Journal of the Learning Sciences **9**, 437 (2000).
[3] A. Traxler, A. Gavrin, and R. Lindell, Physical Review Physics Education Research **14**, 020107 (2018).
[4] K. Neuendorf, *The Content Analysis Guidebook* (Sage Publications, 2002).
[5] L. Rourke, T. Anderson, D. R. Garrison, and W. Archer, International Journal of E-Learning & Distance Education **14**, 50 (1999).
[6] G. Kortemeyer, American Journal of Physics **74**, 526 (2006).
[7] B. De Wever, T. Schellens, M. Valcke, and H. Van Keer, Computers & Education **46**, 6 (2006).
[8] M. Gamer, J. Lemon, I. Fellows, and P. Singh, *irr: Various Coefficients of Interrater Reliability and Agreement* (2012), R package version 0.84.
[9] R. Huang, *RQDA: R-based Qualitative Data Analysis* (2017), R package version 0.3-0.